UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Identificação e caracterização de pequenos RNAs não codificantes e genes alvos envolvidos em estresse abiótico (seca e salinidade) em *Eugenia uniflora* L. (Myrtaceae)

María Eguiluz Moya

Porto Alegre

2018

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Identificação e caracterização de pequenos RNAs não codificantes e genes alvos envolvidos em estresse abiótico (seca e salinidade) em *Eugenia uniflora* **L. (Myrtaceae)**

María Eguiluz Moya

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal do Rio Grande do Sul como requisito parcial para obtenção do grau de Doutor em Ciências (Genética e Biologia Molecular).

Orientador: Prof. Dr. Rogerio Margis

Coorientador: Dr. Frank Guzman Escudero

Porto Alegre, abril de 2018

INSTITUIÇÕES E FONTES FINANCIADORAS

Este trabalho foi realizado no Laboratório de Genômica e Populações de Plantas, Centro de Biotecnologia e Departamento de Biofísica da Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, com apoio financeiro da FAPERGS, MCTIC e CNPq. A doutoranda obteve bolsa de estudos de Innovate Peru, Instituição peruana de Inovação (48 meses).

À minha família

AGRADECIMIENTOS

Aos meus pais, por sempre me apoiar em todas minhas decisões e acreditar em mim.

A minha irmã, pela força, paciência e conselhos fornecidos durante o desenvolvimento da tese.

A Lucila, por me acolher na sua casa, me ensinar a cultura brasileira e ser como uma mãe para mim.

Ao meu orientador, Rogerio, pela oportunidade de trabalhar sob sua orientação, pela confiança, e principalmente paciência e compreensão no desenvolvimento deste trabalho.

Ao meu coorientador, pela orientação, ajuda e amizade brindada neste trabalho.

Aos colegas do laboratório LGPP, Isabel, Erika, Henrique, Pabulo, e Débora pelo apoio, amizade, parceria nestes anos de doutorado. Especialmente a Guilherme Cordenonsi e a Nurevey por me ajudar com o português.

Aos amigos peruanos e brasileiros, por todas as conversas pelo Skype e todas as viagens pelo Rio Grande do Sul.

Ao governo peruano através de *Innovate Peru* pela concessão da bolsa.

Aos componentes da banca, por aceitarem avaliar e contribuir na finalização deste trabalho.

Ao Elmo, pelo auxílio e presteza em todos os momentos.

SUMÁRIO

ABREVIATURAS

AF—Floresta Atlântica

AGO—Argonauta

Ala—Alanina

ARF—fator de transcrição responsivo à auxina

DCL1—proteína tipo Dicer 1

EST—sequências de expressão etiquetadas do inglês Expressed Sequence Tag

Gly—Glicina

HEN1—potenciador Hua1

miRNAs—microRNAs

mRNAs- RNAs mensageiros

NPR1—receptor peptídico natriurético A

nt—nucleotídeos

PARE-Seq—Análise em paralelo das sequências das extremidades dos RNAs

PEG—Polietileno glicol

PolII—RNA polimerase II

Pre-miRNA—precursor do miRNA

Pri-miRNAs—microRNAs primários

RISC—complexo de silenciamento induzido por RNA

RT-qPCR—Transcrição reversa seguida do PCR quantitativo.

Ser—Serina

sncRNAs—pequenos RNAs não codificantes

sRNAs—pequenos RNAs

Thr—Treonina

tRFs—fragmentos derivados do RNA transportador

tRNA—RNA transportador

Tyr—Tirosina

Val—Valina

RESUMO

As plantas, por serem organismos sésseis, enfrentam persistentemente perturbações ambientais adversas denominadas estresses abióticos, sendo as mais importantes, a seca, a salinidade do solo, as temperaturas extremas e a presença de metais pesados. Em resposta, as plantas desenvolveram mecanismos de tolerância, resistência e prevenção para minimizar a influência do estresse, utilizando estratégias de curto prazo para readaptar rápida e eficientemente seu metabolismo. Neste sentido, os pequenos RNAs não codificantes (sncRNAs) são fortes candidatos para realizar este tipo de regulação. Através do sequenciamento de nova geração revelou-se o papel dos sncRNAs na regulação da expressão gênica em nível transcricional e pós-transcricional. Dentre os sncRNAs, os microRNAs (miRNAs) são os mais conhecidos e os fragmentos derivados dos RNAs transportadores (tRFs) são os mais novos e com maiores perspectivas de descobertas futuras. Os miRNAs desempenham papéis regulatórios essenciais tanto no crescimento das plantas quanto no desenvolvimento e resposta ao estresse, enquanto os tRFs, em sua maioria, têm sido associados a respostas de estresse.

Eugenia uniflora L., “pitanga” ou a cereja brasileira é uma árvore frutífera nativa da América do Sul que pertence à família Myrtaceae. Ela cresce em diferentes ambientes; florestas, restingas e ambientes áridos e semi-áridos no nordeste brasileiro, sendo uma espécie versátil em termos de adaptabilidade e que desempenha um papel fundamental na manutenção da vegetação costeira arbustiva. Além disso, é muito conhecida por suas propriedades medicinais que são atribuídas aos metabólitos especializados presentes nas folhas e frutos. *E. uniflora* representa uma fonte fascinante da biodiversidade do germoplasma e tem um grande potencial como fonte de genes para o melhoramento genético. Portanto, a compreensão dos mecanismos que conferem tolerância ao estresse nesta planta é de particular importância. Nesse contexto, o objetivo do presente trabalho é a identificação de sncRNAs (miRNAs e tRFs) por ferramentas de bioinformática e análise do padrão de expressão destes sob condições de estresse abiótico (seca e salinidade), bem como avaliação dos genes envolvidos nesta resposta.

No capítulo 1, bibliotecas de DNA, pequenos RNAs (sRNAs) e RNAseq de folhas foram usadas para identificar pre-miRNAs, miRNAs maduros e potenciais alvos destes miRNAs, respectivamente. A montagem *de novo* do genoma permitiu identificar 38 miRNAs conservados e 28 novos miRNAs. Após a avaliação da expressão destes, 11 conservados, entre eles miR156 e miR170, mostraram variação significativa nas condições de restinga e de estresse induzido por PEG. A maioria deles foram previamente descritos em processos de estresse em outras espécies. 14 novos miRNAs foram avaliados em diferentes tecidos de pitanga mostrando variação significativa no padrão de expressão. Os alvos destes últimos miRNAs foram preditos e validados por RT-qPCR. Eles correspondem a genes de fatores de transcrição e outros genes como transferases ou ATPases e demonstraram o padrão esperado oposto à expressão dos miRNAs.

No capítulo 2, as mesmas bibliotecas foram usadas para identificar tRFs conservados na família das Myrtaceae. Para isso, os tRNAs de *Eucalyptus grandis* e *E. uniflora* foram anotados e os tRNAs comuns foram utilizados para o ancoramento dos sRNAs. 479 tRFs foram identificados em pitanga, na maioria com 18 nucleotídeos (nt). Um conjunto de 11 tRFs conservados em ambas espécies, assim como seus alvos, foram avaliados em condições de estresse salino e seca demonstrando diferenças significativas dependendo do tipo de estresse. Os alvos identificados correspondem a genes previamente descritos como envolvidos em estresse salino e seca para outras espécies.

O presente trabalho apresenta fortes evidências do envolvimento dos miRNAs em processos de desenvolvimento e estresse, assim como dos tRFs na resposta à seca e estresse salino presente em *E. uniflora*. Além disso, os dados produzidos poderão ser utilizados em estudos funcionais mais aprofundados que servirão para melhor compreensão dos mecanismos de tolerância presentes nesta importante planta.

ABSTRACT

Plants being sessile organisms, persistently face adverse environmental perturbations termed as abiotic stresses, most important being drought, soil salinity, extreme temperatures, and heavy metals. They developed several strategies such as tolerance, resistance, and avoidance to minimize stress influence, thus require short-term strategies to quickly and efficiently readapt their metabolism. In this sense, small non coding RNAs are strong candidates to do this kind of fine tune regulation. Next generation sequencing technologies have revealed the key role of these sncRNAs in the transcriptional and post-transcriptional gene-expression regulation. Among the myriad of new sncRNAs, miRNAs are the most known ones and the fragments derived from tRNAs (tRFs) are the newest but with high perspective ones. The miRNAs are endogenous small RNAs that play essential regulatory roles in plant growth, development and stress response. In the case of tRFs, they are mainly involved in stress response.

Eugenia uniflora L., 'pitanga' or Brazilian cherry is a fruit tree native to South America that belongs to Myrtaceae family. It grows in several different harsh environments, including forests, restingas, near the beach, and arid and semiarid environments in the Brazilian northeast. This species is very versatile in terms of adaptability and plays a fundamental role in the maintenance of the shrubby coastal vegetation. However, this species is best-known because its medicinal properties that are attributed to specialized metabolites with known biological activities present in their leaves and fruits. *E. uniflora* is a fascinating reservoir of germplasm biodiversity and has great potential as a source of genes for plant breeding. Therefore, understanding the mechanisms conferring stress tolerance will be very useful. In this sense, the objective of this work is to identify sncRNAs (miRNAs and tRFs) by bioinformatic tools and to analyze their expression pattern under stress conditions as well as the genes involved in that response.

In chapter 1, DNA, small RNA (sRNA) and RNAseq libraries from leaves were used to identify pre-miRNAs, mature miRNAs and potential targets of these miRNAs, respectively. *De novo* assembly of the genome identified 38 conserved miRNAs and 28 novel miRNAs. After evaluating their expression pattern, 11

conserved miRNAs, including miR156 and miR170, showed significant variation in the natural (restinga habitat) and PEG induced stress. Most of them were previously reported in stress processes. 14 novel miRNAs were evaluated in different tissues of pitanga showing significant variation in the expression pattern. The targets of the last miRNAs were predicted and validated by RT-qPCR. They were transcription factor genes and other genes such as transferases or ATPases and showed the expected opposite pattern to miRNA expression.

In Chapter 2, the same libraries were used to identify conserved tRFs in the Myrtaceae family. To do this, the tRNAs of *Eucalyptus grandis* and *E. uniflora* were annotated and sRNAs mapped into them. 479 tRFs were identified in pitanga with predominance of those with 18 nucleotide length. 11 conserved tRFs in both species, as well as their targets, were evaluated under saline and drought stress conditions showing significant differences depending on the stress type. The targets were genes previously involved in saline and drought stress for other species.

The present work shows strong evidences of the involvement of the miRNAs in the development and stress, as well as the tRFs in the tolerance to drought and saline stress of *E. uniflora*. In addition, the data could be used in more detailed functional studies that will serve to corroborate and better understand the mechanism of tolerance present in this important plant.

1. INTRODUÇÃO

1.1. Floresta Atlântica

A Floresta Atlântica (FA) é um das ecoregiões presentes na América do Sul, sendo considerada a segunda maior floresta tropical deste continente e destaca-se como um dos principais centros de biodiversidade do mundo (Myers et al. 2000). Ela cobre uma área de mais de um milhão de quilômetros quadrados ao longo da costa brasileira, estendendo-se até o leste do Paraguai e nordeste da Argentina (Oliveira-Filho and Fontes 2000; Ribeiro et al. 2009). Conforme o Instituto Brasileiro de Geografia e Estatística (IBGE 1988) a Floresta Atlântica é formada por um conjunto de diferentes formações vegetais: Florestas Ombrófila Densa, Ombrófila Aberta e Mista, Floresta Estacional Decidual e Semidecidual, campos de altitude, Manguezais, Restingas e Dunas. Todas elas estão presentes dentro do bioma FA ou na borda dele. Porém, a diversidade das comunidades de plantas na periferia é menor porque elas são submetidas a condições ambientais adversas mais extremas tais como altas e baixas temperaturas (incluindo congelamento), seca, alagamentos, constantes ventos, alta salinidade e falta de nutrientes. Por exemplo, nas restingas, dispersas em toda a costa brasileira, as plantas estão sujeitas a salinidade atmosférica, alta radiação solar, oligotrofia do solo e baixa disponibilidade de água (Scarano et al. 2001). Dentre todo o conjunto de famílias de plantas, Myrtaceae, depois das Leguminosae, é a família das lenhosas mais rica (Oliveira-Filho and Fontes 2000) ou a segunda mais rica (Stehmann et al. 2009) em espécies presente na FA e na sua borda.

1.2. Familia Myrtaceae

Myrtaceae é a oitava maior família de plantas com flores que inclui aproximadamente 142 gêneros e cerca de 55000 espécies. A sua distribuição fica concentrada na Austrália, no Sudeste Asiático e na região Neotropical, além de uma pequena representação na África. A família domina vários tipos de vegetação na América do Sul através de uma variedade de ecótipos (Wilson et al. 2001).

Durante muito tempo, a família foi dividida em duas subfamílias: Myrtoideae, composta de frutos carnosos, folhas opostas e distribuição

pantropical; e Leptospermoideae, caracterizada por frutos secos, folhas alternas e distribuição na Oceania. Diversos trabalhos mostraram que estes grupos não são monofiléticos e baseado em estudos filogenéticos sugerem a reorganização da família, considerando como subfamílias Psiloxyloideae e Myrtoideae contendo 2 e 15 tribos, respectivamente (Wilson et al. 2001; Wilson et al. 2005).

Dentro de Myrtoideae encontra-se duas tribos muito importantes para este estudo: Eucalypteae e Myrteae. A primeira constituída por gêneros como *Eucalyptus*, *Corymbia* e *Angophora* e a Myrteae que engloba todas as espécies neotropicais com exceção do gênero *Tepualia* da família Myrtaceae. A tribo Myrteae apresenta distribuição pantropical (Govaerts et al. 2015), mas sua diversidade é concentrada na América tropical, principalmente na porção leste (Floresta Atlântica), no Planalto das Guianas e no Caribe (Mcvaugh 1968).

1.3. *Eugenia uniflora* L.

Eugenia uniflora L., 'pitanga' ou cereja brasileira faz parte da tribo Myrteae, do gênero *Eugenia*, o maior dentre as Myrtaceae neotropicais, com mais de 1050 espécies (Mazine et al. 2014). Ela cresce em uma variedade de regiões fitogeográficas na AF, incluindo a Floresta tropical, a floresta semidecidual (Oliveira-Filho and Fontes 2000), pampa brasileira (Roesch et al. 2009) e restinga (Scarano 2002), desde o Nordeste até a região sul do Brasil, norte da Argentina e Uruguai. Esta espécie consegue crescer em todas essas regiões sendo muito versátil em termos de adaptabilidade. Por exemplo, é um arbusto ou uma pequena árvore na vegetação arenosa da planície costeira perto do oceano no Sudeste e no nordeste do Brasil; ou uma árvore na parte sul da FA (Oliveira-Filho and Fontes 2000; Almeida et al. 2012; Lucas and Bünger 2015) (Figura 1).

Figura 1. *Eugenia uniflora* como arbusto em ambientes de restinga na praia de Grumari, Rio de Janeiro (A) e como uma árvore em Porto Alegre no Rio Grande do Sul, sudeste do Brazil (B).

Alguns trabalhos sobre diversidade genética em populações dessa espécie utilizando marcadores moleculares mostraram que a maior percentagem de diversidade genética foi observada dentro das populações (Margis et al. 2002; Salgueiro et al. 2004). Um estudo filogeográfico de *E. uniflora* ao longo de toda a sua distribuição corroboram esses resultados e revelaram uma alta variabilidade nas populações dessa espécie no extremo sul de sua distribuição, enquanto uma menor variabilidade foi detectada em populações nas regiões sudeste e nordeste (Turchetto-Zolet et al. 2016).

Pelo mesmo motivo que a pitanga consegue se desenvolver em ecossistemas muitas vezes hostis, estudos indicam *E. uniflora* como uma planta tolerante ao estresse. Ela consegue reagir e se adaptar rapidamente devido à ativação de uma série de mecanismos bioquímicos e fisiológicos, tais como, diminuição da atividade fotossintética, incremento da acumulação de osmólitos (como a prolina) ou pelo ativação de enzimas como a superóxido dismutase e a catalase (Toscano et al. 2016).

Além disso, *E. uniflora* possui frutos comestíveis que se caracterizam pelo baixo conteúdo de lipídeos e calorias, e pela presença de substâncias biologicamente ativas, principalmente compostos fenólicos e carotenóides, que fazem dela uma fonte de antioxidantes (Spada et al. 2008). Além dos frutos, as folhas da pitanga também são usadas na medicina popular como infusões no tratamento da febre, reumatismo, doenças estomacais e hipertensão (Lim 2012) (Figura 2).

Figura 2. Frutos da pitanga crescida em restinga do Rio de Janeiro**.** Fotos do professor Fabiano Salgueiro.

1.4. Pequenos RNAs não codificantes (sncRNAs)

A regulação pós-transcricional representa uma rede integrada de um conjunto de RNAs, dentre as quais os RNA regulatórios desempenham um papel importante na decodificação e regulação génica. O aumento das abordagens de sequenciamento de nova geração tem amplamente elucidado várias classes de RNAs reguladores entre eles os categorizados como pequenos RNAs não codificantes (sncRNAs) (Heo et al. 2013).

Em várias espécies de plantas a maioria dos trabalhos com sncRNAs estão focados na classe mais abundante deles os miRNAs (Zhang et al. 2006). Eles coordenam muitas atividades regulatórias e redes de interação, as quais demonstram o papel fundamental que eles desempenham na compreensão da genômica funcional e desenvolvimento das plantas (Meng et al. 2011; Li and Zhang 2016). Com o avanço das tecnologias de sequenciamento, uma nova classe de sncRNAs tem aparecido, os fragmentos derivados de RNAs transportadores (tRFs). Eles ganharam importância substancial na genômica de plantas por apresentarem interações canônicas com a proteína AGO1, assim como os miRNAs (Alves et al. 2017; Martinez et al. 2017), e também por estarem envolvidos em processos de estresse.

a) microRNAs (miRNAs)

Os miRNAs maduros são sequências pequenas de RNA de cadeia simples com 18-24 nucleotídeos de comprimento que se ligam por complementaridade de bases ao seus RNA mensageiros (mRNA) alvo. Eles

podem regular a expressão gênica no nível pós-transcricional através dá clivagem ou inibição da tradução de seus mRNAs alvos e no nível transcricional através da remodelação da cromatina e/ou metilação do DNA. Geralmente, a biogênese dos miRNAs envolve várias etapas inter dependentes, que incluem a transcrição primária dos miRNAs (pri-miRNAs), seu processamento e modificações, até o carregamento no complexo de silenciamento induzido por RNA (RISC) (Li and Zhang 2016) (Figura 3).

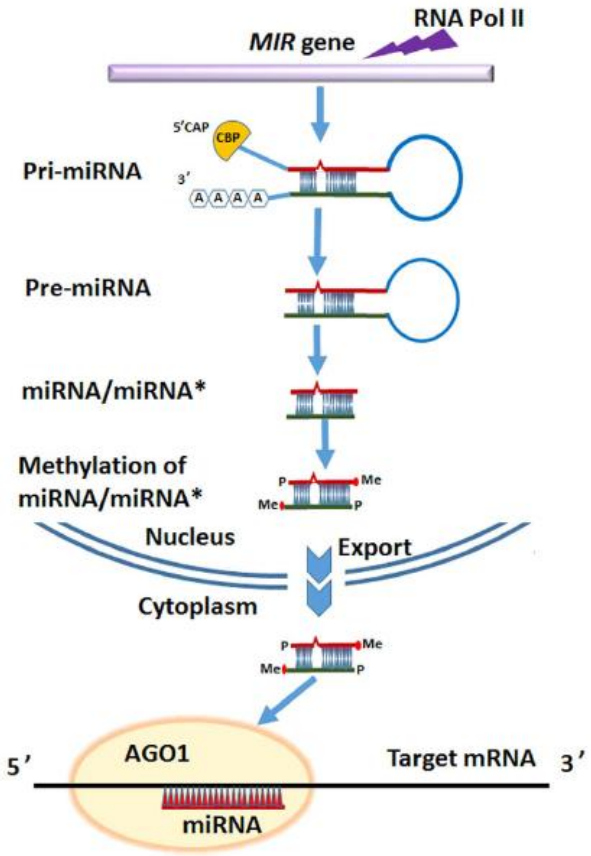


Figura 3. Representação gráfica do processo de biogêneses dos miRNAs. Figura tomada de Li and Zhang (2016). "*MicroRNA in control of plant development*"

A transcrição dos pri-miRNAs de plantas é semelhante ao processo dos genes codificantes. A maioria deles são transcritos a partir de suas próprias unidades transcricionais denominadas genes MIR, cujas sequências genômicas localizam-se geralmente nas regiões intergênicas, que possuem seus próprios promotores e padrões de regulação independentes (Nozawa et al. 2012). Nas plantas, o ativador transcricional dependente de DNA, a RNA polimerase II (Pol II), é aquela que transcreve, na maioria das vezes, os genes MIR produzindo os miRNAs primários (pri-miRNAs). A eles se adiciona o *cap* na extremidade 5' e

uma cauda de adeninas na extremidade 3’ antes do processamento posterior (Lee et al. 2004).

Os pri-miRNAs possuem uma estrutura de grampo imperfeita necessária para direcionar a DICER-LIKE1 (DCL1) no processo de clivagem e com isso gerar o precursor dos miRNAs (pre-miRNA). Esta estrutura é processada depois pela mesma DCL1 gerando o duplex miRNA/miRNA*. Para aumentar a estabilidade a extremidade 3’ é metilada pela RNA metiltransferase hua enhancer 1 (HEN1) no núcleo. Em *Arabidopsis*, a proteína argonauta 1 (AGO1), que possui atividade de endonuclease, é a encarregada de recrutar o miRNA para formar o complexo de silenciamento induzido por RNA (RISC). Neste complexo, uma fita simples do pequeno RNA maduro funciona como guia para a posterior degradação ou inibição do mRNA alvo (Bologna and Voinnet 2014). Muitos indicam a funcionalidade de ambos miRNAs do duplex, portanto o par miRNA/miRNA* foi renomeado para miR-5p e miR-3p (Desvignes et al. 2015).

Os miRNAs tem algumas características principais que ajudam no processo da sua identificação, entre elas: (1) todos os miRNAs são sncRNAs, geralmente ~21-24 nucleotídeos (nt) de comprimento em plantas (2) todos os precursores de miRNAs formam uma estrutura de grampo cujo rearranjo tridimensional mais provável é o de menor energia livre. Além disso, miR-5p e miR-3p são derivados de braços opostos neste grampo, de modo que eles devem formar um duplex com dois nucleotídeos 3’ sobressalentes no final. O extenso pareamento de bases entre eles não permite mais que 3 ou 4 bases não pareadas e tem pouca presença de alças dentro do duplex. As estruturas de grampo podem ser preditas por programas computacionais tais como o RNAfold (Hofacker 2003) (3) Muitos miRNAs maduros são conservados evolutivamente (Bartel 2004). No entanto, algumas dessas características não são únicas para os miRNAs, por isso, é importante incluir alguns critérios no momento da identificação de gene candidatos de miRNAs, especialmente quando novos miRNAs são reportados (Axtell and Meyers 2018).

Os miRNAs podem ser identificados por quatro abordagens diferentes: o *screening* genético (Lee et al. 1993; Wightman et al. 1993), a clonagem direta após o isolamento dos sncRNAs (Lu et al. 2005), a análise de sequências de expressão etiquetadas (*Expressed Sequence Tag* ou ESTs) (Zhang et al. 2005)

e a análise de dados provenientes do sequenciamento de nova geração. Esta última abordagem baseia-se na análises de genomas e usos de programas computacionais como o miRDeep2 (Friedländer et al. 2012) ou miRPREFeR (Lei and Sun 2014) para predizer miRNAs. Todos eles já demonstraram sucesso na identificação de genes de miRNAs em *Arabidopsis* (Yang *et al.* 2011), café (Loss-Morais et al. 2014), arroz (Yang and Li 2012; Wen et al. 2016), tomate (Liu et al. 2017), *Catharanthus roseus* (Shen et al. 2017) entre outros. O passo seguinte é confirmar os níveis de expressão destes miRNAs candidatos preditos pelos programas de bioinformática, para isso podem-se usar diferentes metodologias como o *Northern Blot*, a PCR quantitativa seguida da transcrição reversa (RT-qPCR) ou os microarranjos de miRNAs. Porém, na atualidade, é a RT-qPCR a técnica mais usada nos processos de quantificação e validação da expressão de miRNAs em plantas (Zhang and Wang 2015).

A técnica da PCR quantitativo em tempo real é o procedimento preferencial na quantificação da expressão gênica. No caso dos miRNAs foi proposto uma modificação no desenho dos iniciadores para conseguir quantificar estes fragmentos pequenos (Chen et al. 2005) (Figura 4).

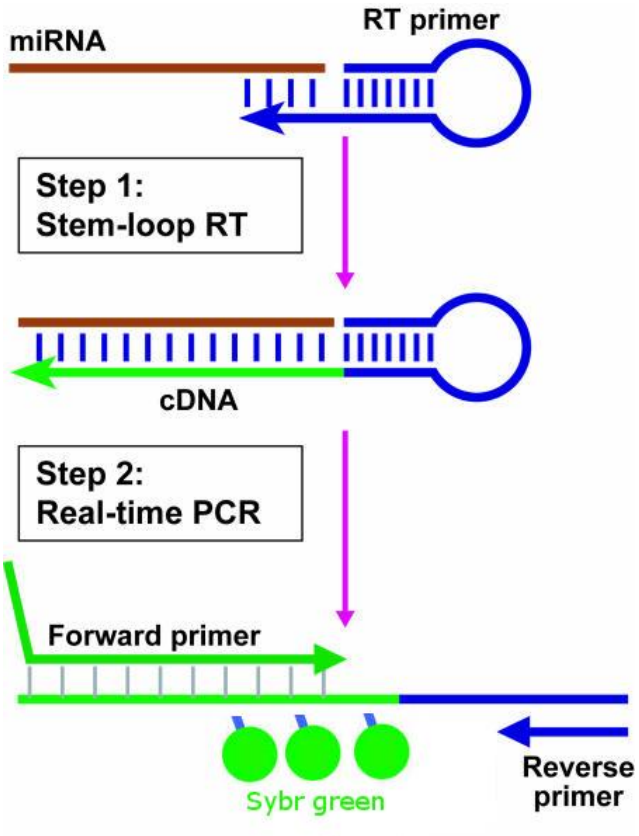


Figura 4. Esquema da metodologia usada na quantificação dos miRNAs. O processo tem dois passos, uma transcrição reversa (RT) e uma PCR quantitativa em tempo real. No primeiro, os iniciadores se ligam à porção 3' das moléculas de mRNA e são transcritos pela transcriptase reversa. Em seguida, o produto RT é quantificado usando PCR TaqMan convencional que inclui o iniciador direto específico do miRNA, o iniciador reverso universal. Figura modificada de Chen et al., (2005). "*Real-time quantification of microRNAs by stem–loop RT–PCR"*

A técnica baseia-se na formação de um *stem-loop* para a transcrição reversa (RT) com ajuda de iniciadores específicos que contém a sequência complementar reversa dos seis nucleotídeos finais do miRNA maduro, assim como nucleotídeos adicionais que formam a estrutura de loop. Estas duas características permitem aumentar a estabilidade, especificidade e sensibilidade desta técnica com respeito a uma PCR quantitativa que usa iniciadores lineares. Depois, da RT vem o processo de amplificação normal com o uso de um iniciador universal que se ancora no loop e um iniciador específico que possui uma sequência parcial do miRNA maduro, não contendo os seus seis últimos nucleotídeos. Inicialmente a técnica foi proposta usando sondas Taq-man, porém o uso de corantes fluorescentes intercalantes como o Sybr-green também tem sido reportado para a quantificação em tempo real (Kulcheski et al. 2011).

Com o surgimento das tecnologias de sequenciamento de nova geração, o número de miRNAs de plantas, identificados e anotados funcionalmente, tem aumentado exponencialmente, levando ao estabelecimento de bancos de dados biológicos que atuam como arquivos de sequências e anotações de miRNAs, tal como o miRBase (Griffiths-Jones et al. 2008). Nesta base de dados, alguns dos genes de miRNAs estão organizados em diferentes grupos, denominados famílias de miRNAs, baseados no RNA maduro e na sequência e/ou estrutura dos pré-miRNAs. As famílias de miRNA são importantes porque sugerem uma sequência ou configuração de estrutura comum neste conjunto de genes o que indicaria uma função similar também (Griffiths-Jones et al. 2008). O conjunto de análises funcionais e experimentais indicam que existem miRNAs e seus alvos conservados desde musgos até eudicotiledôneas. Estes miRNAs conservados são quase idênticos ou tem apenas algumas mudanças de nucleotídeos entre eles. Dentre estes, existe uma pequena porção de miRNAs presentes em várias grandes linhagens de plantas terrestres. No total, na versão atual do miRBase (versão 22) tem 39 famílias de miRNAs presentes em duas ou mais espécies de

plantas filogeneticamente distantes. Os miRNAs conservados desempenham um papel importante na regulação de genes conservados, como a morfologia das folhas e das flores, desenvolvimento, ou a transdução de sinal. Já os miRNAs não conservados pelo contrário, são menos abundantes e podem desempenhar papéis mais específicos em cada espécie de plantas, como a resistência ao estresse ou formação da fibra no caso do algodão por exemplo (Zhang et al. 2013; Zhang 2015).

Os miRNAs regulam na maioria das vezes fatores de transcrição. Eles são indicados como os principais coordenadores no crescimento e desenvolvimento, nas respostas ao estresse e no *crosstalk* em diferentes vias de transdução de sinais nas plantas. Portanto, alterações na expressão deles irão resultar em mudanças significativas para o organismo (Rubio-Somoza and Weigel 2011; Kamthan et al. 2015). Por exemplo, existem miRNAs que atuam como reguladores chaves no desenvolvimento da raiz através da regulação do fator de transcrição dependente de auxina (ARFs) (Khan et al. 2011) e também atuam no crescimento de frutos em *Arabidopsis* (José Ripoll et al. 2015) ou nos processos de indução floral e formação de flores (Hong and Jackson 2015). Itaya et al. (2008) identificaram um grande número de miRNAs espécie-específicos de tomate (*Solanum lycopersicon*) que atuavam no desenvolvimento do fruto (Itaya et al. 2008). Além disso, esses miRNAs já foram identificados como reguladores de genes envolvidos em processos como a assimilação de enxofre e a degradação de proteínas dependentes de ubiquitina (Bonnet et al. 2004). Uma série de miRNAs foram identificados tendo como genes alvos aqueles envolvidos em rotas metabólicas de enchimento de grãos e biossíntese de nutrientes, incluindo o metabolismo de carboidratos e proteínas, transporte celular e transdução de sinais além da sinalização de fito-hormônios. Todos estes estudos propõem o uso dos miRNAs específicos como uma estratégia inovadora e potente para melhorar o crescimento da planta, a biomassa e o rendimento das culturas (Zhang and Wang 2015). Por exemplo, tem um estudo mostrando que a superexpressão do miR156 resulta no aumento do rendimento de biomassa vegetal da grama (*Panicum virgatum*). As plantas transgênicas conseguiram um 58% -101% a mais de rendimento de biomassa vegetal do que as plantas silvestres. Este aumento na produção foi devida à inibição da dormência apical.

Com isso, nas plantas transgênicas geradas teve um aumento da produção de biocombustíveis na fase posterior (Fu et al. 2012).

b) Fragmentos derivados de RNA transportadores (tRFs)

Os RNAs transportadores (tRNAs) são aqueles que fornecem os aminoácidos ao ribossomo para síntese de proteínas. Além desse papel fundamental, eles podem assumir outras funções biológicas interagindo com uma ampla gama de proteínas envolvidas em vias de regulação e sinalização. Eles conseguem cumprir essas funções adicionais porque podem ser clivados por ribonucleases específicas dando origem aos denominados fragmentos derivados de RNAs transportadores (tRFs) (Cole et al. 2009; Soares and Santos 2017).

Como já foi mencionado, eles foram descobertos como consequência das novas tecnologias de sequenciamento. Inicialmente foram identificados como produtos de degradação devido a sua grande abundância nas bibliotecas de sequenciamento de sRNAs. Porém, a abundância de *reads* que mapeiam num domínio específico do tRNA maduro sugeriu que estes tRFs poderiam ser funcionais e não fragmentos gerados aleatoriamente (Lee et al. 2009).

Os tRFs estão universalmente identificados em todo os domínios da árvore da vida. Eles foram descritos em bactérias (Kumar et al. 2014), algas (Åsman et al. 2014), archaea (Gebetsberger et al. 2012), protozoa (Liao et al. 2014), vermes (Cai et al. 2013), plantas (Chen et al. 2011; Alves et al. 2017), leveduras (Bühler et al. 2008) e mamíferos (Kawaji et al. 2008; Liao et al. 2010; Telonis et al. 2015). Com toda essa informação, criou-se uma base de dados de tRFs onde estão depositados 552, 559, 433, 320 e 649 tRFs correspondentes aos humanos, camundongos, *Drosophila*, *S. pombe* e *C. elegans,* respectivamente (Kumar et al. 2015).

Os tRFs são classificados dependendo da posição de onde eles são gerados dentro dos tRNAs. Os dois tipos principais são: os 5'tRFs, também conhecidos como tRF-5 que derivam da clivagem da extremidade 5' do tRNA maduro perto do D-*loop* e os 3'tRFs, ou tRF-3, que são originados pelo processamento da extremidade 3' do tRNA maduro no TψC-*loop* que contém a modificação pós-transcricional dos três nucleotídeos CCA. Existem outros tRFs

também reportados na literatura como os 3'-U-tRFs que derivam do processamento do pré-tRNA e contêm resíduos poli-U na extremidade 3', as metades do tRNA (tRNA *halves*) principalmente envolvidas em estresse e por último os tRFs endógenos (i-tRFs) que derivam da clivagem dos domínios internos dos tRNAs maduros (Figura 5).

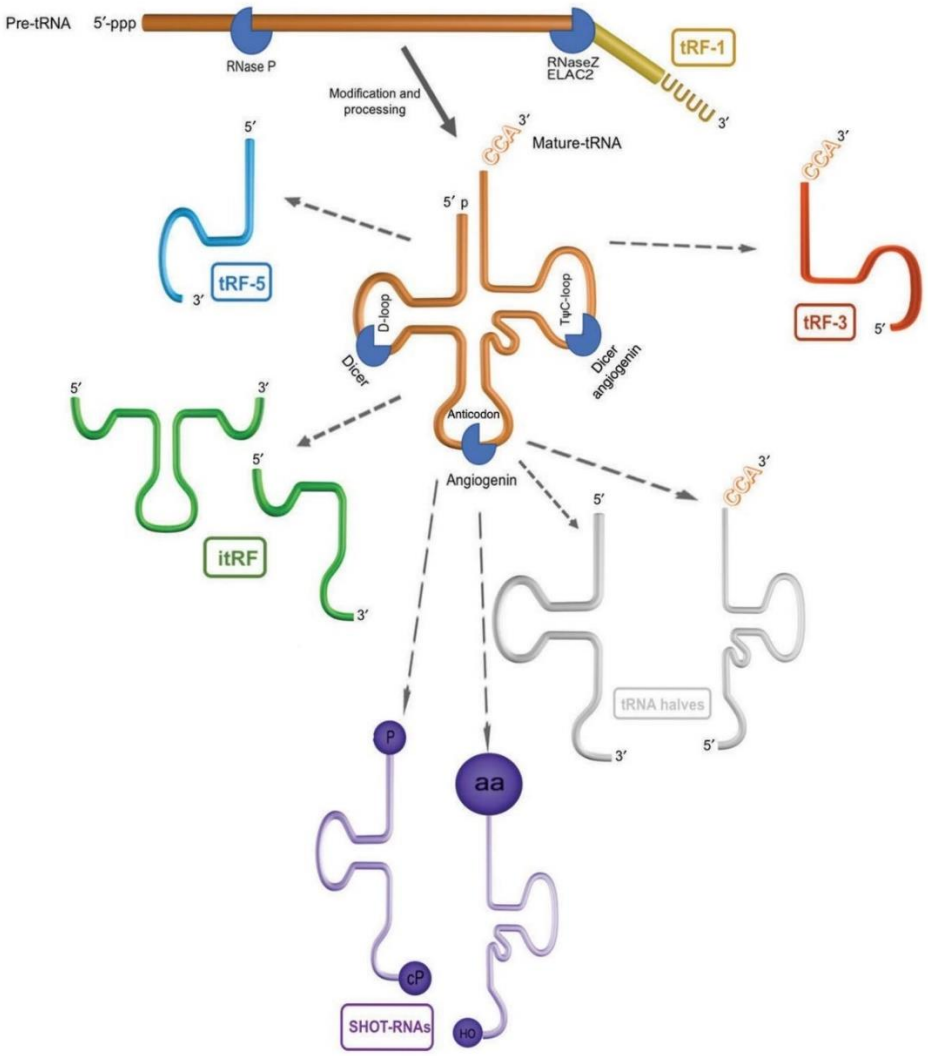


Figura 5. Diferentes tipos de fragmentos derivados de RNA transportadores(tRFs) produzidos a partir do pré-tRNA ou do tRNA maduro. A Rnase P remove a extremidade 5' do pré-tRNA transcrito e a RNase Z (ELAC2) remove a extremidade 3'. Os 3'-U-tRFs ou tRF-1 (amarelo) são produzidos após a clivagem da molécula de pré-tRNA pela RNAseZ (ou ELAC2). Vários tRFs são produzidos por clivagem endonucleolítica dos tRNA maduros. A Dicer e angiogenina, estão envolvidas na biogênese tRFs em vertebrados. tRF-5 ou 5'tRF (azul), tRF-3 ou 3'tRF (vermelho), tRFs endógenos (i-tRFs-verde), RNAs derivados de tRNAs dependentes de hormônios e do sexo (SHOT-RNAs-roxo) e as metades de tRNA (cinza). Figura tomada de Soares e Santos (2017) "*Discovery and function of transfer RNA-derived fragments and their role in disease*"

A biogênese dos tRFs ainda não está totalmente esclarecida. Segundo alguns trabalhos, parece que eles usariam a mesma via de sínteses dos miRNAs. Nos mamíferos, o processo parece ser dependente das proteínas Dicer, porém um estudo usando células HEK293 demonstrou um processamento independente das enzimas Dicer e Dgcr8 e analisando tRFs de *Phytophthora infestans, Drosophila melanogaster,* camundongo e *Schizosaccharomyces pombe* demonstrou-se que a maquinaria de sínteses do miRNA canônico não era necessária (Kumar et al. 2014). Já nas plantas, o cenário é parecido com os mamíferos. Inicialmente eles foram indicados como processados pelas proteínas tipo Dicer (Martinez et al. 2017) devido a observação de uma diminuição significativa na produção dos tRFs em mutantes de *dcl1* e *ago1*. Porém, no mesmo ano apareceu outro trabalho analisando tRFs em *A. thaliana*, *Oryza sativa* e *Physcomitrella patens* que propôs um mecanismo de processamento independente das Dicer (Alves et al. 2017). Uma das razões da hipervariabilidade no mecanismo de geração dos tRFs é o fato que pode ser o resultado de mecanismos diferentes ou podem haver fatores tecido-específicos que determinam o tipo de tRF produzido. Outra razão pode ser que sejam um conjunto de endonucleases as responsáveis pela produção dos tRFs e não uma enzima só (Sobala and Hutvagner 2013), tendo em vista que os tRFs estão restritos ao citoplasma e muitos deles são produzidos por outras endonucleases como as angiogeninas (Ivanov et al. 2011). No entanto, organismos como bactérias e arqueias conseguem gerar tRFs ainda não tendo o mecanismo de sínteses de miRNAs canônico. Reforçando isso aí, numerosos estudos mostraram que os tRFs podem ser produzidos na ausência de muitas destas ribonucleases convencionais, o qual indicaria um sistema de regulação mais antigo que não depende de vias de silenciamento gênico convencional (Keam and Hutvagner 2015). É importante mencionar que embora os procariontes não têm muitas das enzimas envolvidas no processamento dos miRNAs, eles expressam muitas endonucleases dependentes do isotipo do tRNA (Ogawa et al. 1999; Tomita et al. 2000).

Os tRFs apresentam muitas características que ajudam a reforçar a ideia que eles não são produtos de uma clivagem inespecífica das endonucleases. Entre as quais: (1) numerosos trabalhos que demonstram que o processamento

do tRNA em tRFs é notavelmente sítio-específico, gerando tRFs com comprimentos definidos entre os diferentes tipos de células. Assim por exemplo em *A. thaliana* os tRFs de 19 nt são os mais abundantes, já nos cereais é um pouco diferente, em *O. sativa* o mais abundante corresponde ao 5'tRF de 25 nt e em *Triticum aestivum* é de 21 nt. Este último é muito importante porque afeta a associação deles com as proteínas AGOs que ao mesmo tempo determinam a eficiência na clivagem e na acessibilidade de seus genes alvos. Alves et al. (2017) demonstraram a associação dos tRFs de 19 e 20 nt de comprimento com AGO2 e AGO5, enquanto com AGO4 foi encontrado uma maior abundância de fragmentos de 19, 24 e 25 nt, respectivamente. (2) a expressão de tRF não se correlaciona com a abundância de seus respectivos tRNAs precursores, com a exceção dos identificados em *Tetrahymena* (Couvillion et al. 2010). A geração de tRFs parece estar restrita a isotipos específicos de tRNAs, em alguns casos determinando o tipo de endonuclease envolvida, sugerindo assim que a seleção e processamento do isotipo do tRNA não são aleatórios. (3) os tRFs exibem características de moléculas reguladoras funcionais. Eles conseguem inibir a transcrição de seus mRNAs alvos de um jeito similar aos miRNA porém a presença de Guanina na extremidade 5' dos tRF comparado com o Uracila ou Adenina presente nos miRNAs indica uma biogênese diferente (Loss-Morais et al. 2013). Essas observações são semelhantes àquelas observadas no fungo patogênico *Phytophthora sojae* (Wang et al. 2016a), sugerindo um padrão de conservação desses nucleotídeos na extremidade 5'. Uma análise interessante revelou que os tRFs podem clivar seus alvos em múltiplos locais em comparação com miRNAs que tem locais de ligação únicos (Wang et al. 2016b). No entanto, usando dados de *PARE-seq*, os tRFs identificados a partir de pólen possuem um sítio específico de clivagem como o que acontece nos miRNAs (Martinez et al. 2017). Estudos recentes utilizando a análise da composição de nucleotídeos através dos sítios de clivagem revelaram um enriquecimento de uracilas em todo o local de clivagem para 5' e 3'tRFs (Wang et al. 2016b). Apesar, das controvérsias ainda presentes, a conservação do local da clivagem fornece suporte de que a origem desses tRFs não é uma digestão exonucleolítica aleatória dos precursores de tRNAs (Alves et al. 2017).

Muitos estudos indicam que os tRFs estão envolvidos na inibição da tradução das proteínas e na repressão transcricional dos seus genes alvos. Na arqueobacteria *Haloferax volcanii*, um 5'tRF específico de 26 nt se une com a subunidade menor do RNA ribossomal reduzindo a sínteses das proteínas e interferindo na atividade da peptidil transferase (Gebetsberger et al. 2012). Nos humanos, o processo está amplamente estudado e é demonstrado a interação com outros sRNAs como miRNAs e pequenos RNAs de interferência (Sobala and Hutvagner 2013). Já nas plantas o cenário é um pouco diferente tendo poucos trabalhos que indiquem o papel dos tRFs. Em *Cucurbita maxima,* tRFs específicos de floema foram detectados por Northern blot e demonstrou-se que interferem com a atividade ribosomal e reprimem a tradução (Zhang et al. 2009). Em *O. sativa*, trabalhando com calos embrionários, reportou-se pela primeira vez a identificação de tRFs entre 20-22 nt em bibliotecas de sRNAs de meristema e a regulação diferencial de 5'tRF AlaAGC e ProCGG em calos e folhas (Chen et al. 2011), o que seria confirmado por trabalhos recentes em *A. thaliana* abordando o tema da existência de tRFs tecido-específicos (Alves et al. 2017). Recentemente, foi identificado uma acumulação de tRFs em pólen que regulam a estabilidade genômica através da interação com transposons em *A. thaliana* (Martinez et al. 2017).

1.5. Os sncRNAs na regulação do estresse

a) miRNAs

Existem muitos estudos que demonstram o padrão de expressão aberrante dos miRNAs sob diferentes tipos de estresse como a salinidade (Xie et al. 2015), deficiência de nutrientes (Liang et al. 2012), radiação UV-B (Wang et al. 2013), calor (Goswami et al. 2014) e estresse por metais pesados (Gupta et al. 2014). Porém, o foco será em seca por ser o estresse testado experimentalmente no nosso trabalho com miRNAs e devido à grande quantidade de estudos de miRNAs em diferentes tipos de estresse. Nesse sentido, existem vários estudos demonstrando a variação na expressão de miRNAs durante as condições de déficit hídrico em espécies de plantas, como feijão de corda (Barrera-Figueroa et al. 2011), a soja (Kulcheski et al. 2011) ou trigo (Kantar et al. 2011) dentre outras. As respostas das plantas variam dependendo da espécie, do habitat ou da família a qual ela pertence, resultando

em diferentes padrões de expressão e acumulação de miRNAs. Entretanto, alguns dos miRNAs podem compartilhar o mesmo padrão de expressão. As diferenças podem ser identificadas pelos genótipos que existem dentro da mesma espécie, como o caso de soja que apresenta dois genótipos diferentes (sensíveis e tolerantes à seca) esperando um padrão diferente na expressão dos miRNAs.

É importante destacar que alguns miRNAs compartilham o mesmo padrão de expressão; por exemplo, o miR474 que é regulado positivamente no milho e no trigo sob déficit de água ou o miR393 que também é regulado positivamente em condições de seca em *Oryza, Arabidopsis, Medicago truncatula* e *Phaseolus vulgaris* (Liu et al. 2008; Arenas-Huertero et al. 2009; Jian et al. 2010; Trindade et al. 2010). As abordagens de pesquisa usando genomas inteiros permitem identificar uma maior quantidade de miRNAs candidatos, como por exemplo, no caso de arroz onde foram identificados 30 famílias de miRNAs, 16 delas sendo negativamente regulados (miR156, miR159, miR168, miR170, miR171, miR172, miR319, miR396, miR397, miR408, miR529, miR896, miR1030, miR1035, miR1050, miR1088 e miR1126) e 14 famílias sendo regulados positivamente (miR159, miR169, miR171, miR319, miR395, miR474, miR845, miR851, miR854, miR896, miR901, miR903, miR1026 e miR1125) sob o estresse por seca; surpreendentemente 9 miRNAs (miR156, miR168, miR170, miR171, miR172, miR319, miR396, miR397, e miR408) apresentaram padrões de expressão opostos aos descritos anteriormente em *Arabidopsis* (Zhou et al. 2010).

b) tRFs

Embora tenham sido descobertos somente na última década, existem vários estudos que falam do potencial papel dos tRFs sob estresse abiótico e biótico. Num dos primeiros trabalhos, foi demonstrado que o 5'tRF GlyTCC é regulado positivamente nas raízes de *A. thaliana* em condições de deficiência de fósforo. Assim, esse estudo revelou a expressão espaço temporal dos tRFs e seu possível envolvimento na resposta ao estresse. No mesmo trabalho quando a origem desses RNAs foi analisada, não foi encontrada correlação dos tRFs com modificações no uso de códons na planta (Hsieh et al. 2009).

As novas tecnologias de sequenciamento também forneceram evidências de um novo subconjunto de pequenos RNAs derivados do genoma do cloroplasto (csRNAs) analisados em repolho chinês. Esses csRNAs incluem aqueles que são derivados de tRNAs. Wang et al. descobriu que a maioria destes últimos ancoravam na extremidade 5' da molécula. Embora eles tenham apresentado apenas uma pequena redução em sua abundância em plântulas expostas ao calor, com relação ao comprimento verificou-se uma alteração em resposta ao estresse térmico. Portanto, eles poderiam desempenhar um papel semelhante as metades de tRNAs previamente identificadas alteradas em diferentes condições de estresse (Wang et al. 2011b).

Um perfil de expressão completo foi estudado em *Hordeum vulgare* em condições de deficiência de fósforo. Hackenberg et al., usando sequenciamento de nova geração, demonstrou que 56 dos 61 tRNAs geraram tRFs tanto na condição de estresse quanto no controle. Seis deles foram regulados positivamente enquanto que quatro deles foram regulados negativamente nas plântulas com deficiência de fósforo. Também foram identificados os 5'tRF GlyTCC e 5'tRF AlaAGC como os mais abundantes. Este estudo reforçou a ideia que o fosforo tem um forte impacto no processamento dos tRNAs (Hackenberg et al. 2013).

Em 2013, Loss-Morais et al. analisou todos os dados de sequenciamento que existiam na época em *A. thaliana*. O objetivo foi caracterizar os tRFs, analisar seus padrões de acumulação sob estresses abióticos (seca, frio e sal) e bióticos (com *Pseudomona syringae*), e sua associação com proteínas AGOs. Foram examinadas 34 bibliotecas de sRNAs, incluindo 25 bibliotecas imunoprecipitadas com anti-AGOs, e foi possível encontrar tRFs tanto em bibliotecas com AGO1 quanto AGO2, 4 e 7. Os 5'tRFs e 3'tRFs foram associados com AGOs, obtendo os mesmos resultados que os encontrados em mamíferos (Haussecker et al. 2010). Curiosamente, também foi possível detectar-se os i-tRFs, mas os mais abundantes e mais diversos foram os 5'tRFs (Loss-Morais et al. 2013).

O potencial papel da clivagem dos tRNAs como mecanismo de defesa ao estresse térmico em poliplóides como *Triticum aestivum* foi demonstrado por Wang et al. 2016. Neste estudo observou-se um aumento na produção de 5'tRFs

nas plântulas expostas a altas temperaturas, sendo os 5'tRFs ValCAC, ThrUGU, TyrGUA e SerCGA os mais abundantes nestas condições e também no estresse osmótico (Wang et al. 2016b).

Nas plantas, a imunidade associada aos patógenos é controlada através dos padrões moleculares associados a microrganismos ou patógenos (MAPS ou PAMPs), que sob infeção patogênica desencadeiam uma resposta de defesa através da ativação da síntese de genes de defesa. Asha e Soniya (2016) demonstraram que os tRFs estão envolvidos na regulação da expressão dos genes de defesa em resposta à infecção por *Phytophthora capsici* em pimenta preta (*Piper nigrum* L.). As análises revelaram a predominância de 5' tRFs da pimenta nas folhas e nas raízes infectadas tendo como alvos os genes de defesa o que indicaria seu papel na resposta ao estresse. A clivagem do mRNA de uma proteína relacionada com a patogênese (NPR1) indicada como alvo do 5′AlaCGC tRF foi testada e apresentou uma diminuição na sua expressão. Isso prova o papel dos tRFs neste tipo de infeção (Asha and Soniya 2016).

Além de demonstrar que os tRFs tem expressão espaço temporal nas plantas, analisando *A. thaliana*, *P. patents* e *O. sativa*, Alves et al. (2017) avaliou o padrão de expressão dos tRFs em condições de estresse oxidativo. Foi confirmada a predominância dos 5'tRFs sob estas condições e identificaram os 5'tRF ArgTCG e 3'tRF TryGTA como os mais abundantes (Alves et al. 2017).

Ambos tipos de sncRNAs com função reguladora em situações de estresse ainda não foram analisados na família Myrtaceae. Apesar da existência de estudos de identificação de miRNAs em *E. uniflora* (Guzman et al. 2012) e em *E. grandis* (Wang et al. 2011a; Levy et al. 2014; Pappas et al. 2015), eles não foram avaliados sob condições de estresse. No caso dos tRFs, não há nenhum estudo em espécies de Myrtaceae.

2. OBJETIVOS

Identificar e caracterizar pequenos RNAs não codificantes e genes codificantes envolvidos no estresse abiótico (salinidade e seca) em *Eugenia uniflora* L. (Myrtaceae).

Objetivos específicos

- Identificar miRNAs novos e conservados usando o primeiro rascunho do genoma de *E. uniflora;*
- Caracterizar o padrão de expressão dos miRNAs conservados em pitangas crescidas em restinga e sob condições de déficit hídrico induzidas por PEG;
- Caracterizar o padrão de expressão dos miRNAs novos, assim como de seus alvos, em diferentes tecidos de *E. uniflora;*
- Identificar tRFs conservados em Myrtaceae usando bibliotecas de pequenos RNAs em *E. uniflora* e *E. grandis*;
- Caracterizar o padrão de expressão desses tRFs, assim como de seus alvos potenciais, sob condições de seca (PEG) e salino (NaCl) induzidos em ambiente controlado.

3. DISCUSSÃO E CONSIDERAÇÕES FINAIS

E. uniflora, a espécie de estudo neste trabalho, é categorizada como uma planta tolerante ao estresse abiótico devido à capacidade de sobreviver em ambientes contrastantes da Floresta Atlântica. Um dos habitats dela são os ambientes de restinga distribuídos por toda a costa do Brasil que têm características de alta salinidade, alta radiação solar, oligotrofia do solo e baixa disponibilidade de água (Scarano 2002). Uma das muitas estratégias que a pitanga poderia estar usando para regular seu metabolismo e assim conseguir se adaptar a esses ambientes seria pela produção de sncRNAs. Nos últimos anos, eles têm sido estudados e indicados como responsáveis pelos mecanismos de regulação envolvidos no desenvolvimento, morfologia e vias de transdução de sinais, assim como nas respostas das plantas ao estresse. Nesse sentido, as moléculas mais conhecidas e amplamente estudadas são os miRNAs. Estes sncRNAs tem um papel fundamental no crescimento, desenvolvimento e na tolerância da planta ao estresse. Além dos miRNAs e como consequência das análises de bibliotecas de pequenos RNAs (sRNAs) geradas por tecnologias de nova geração, há alguns anos tem se reportado a existência de sncRNAs derivados da clivagem dos RNAs transportadores denominados tRFs. Eles estão envolvidos principalmente em estresse. Ambos tipos de sncRNAs tem uma função de regulação pos-transcricional inibindo a tradução de seus genes alvos (Bartel 2004; Keam and Hutvagner 2015).

Neste contexto, o presente trabalho propõe aos miRNAs e tRFs como agentes de regulação envolvidos na tolerância ao estresse presente naturalmente em *E. uniflora* e busca contribuir com um melhor entendimento da fisiologia desta espécie a nível gênico. O trabalho foi dividido em dois artigos: (i) o primeiro focou-se na identificação de miRNAs novos e conservados usando o primeiro rascunho do genoma de *E. uniflora*, e a avaliação da expressão deles nos tecidos da planta assim como em condições naturais de estresse (em restinga) e seca induzida (ii) o segundo artigo buscou identificar e caracterizar os tRFs conservados entre *E. uniflora* e *E. grandis* para depois avaliar a expressão deles e seus alvos em condições de seca e salinidade.

No primeiro artigo, conseguiu-se identificar 38 miRNAs conservados e 28 novos miRNAs usando o primeiro rascunho do genoma de *E. uniflora* e as

bibliotecas de sRNAs. Os dados de sequenciamento de nova geração permitiram identificar miRNAs expressos em baixas quantidades com alta sensibilidade e em larga escala (Ma et al. 2015). O número de miRNAs identificados poderia ter sido maior, mas foi reduzido devido a utilização de parâmetros mais estringentes e rigorosos, conforme recomendados na literatura (Meyers et al. 2008). Apesar do número de candidatos a miRNAs identificados terem sido consideravelmente maiores àqueles obtidos previamente por Guzman et al. (2012), foi possível encontrar alguns miRNAs compartilhados por ambas abordagens.

Os miRNAs são importantes reguladores de processos metabólicos, participando na regulação de genes de resposta ao estresse. Nesse sentido, o padrão de expressão de miRNAs conservados foi avaliado em duas condições de estresse diferentes. A primeira foram amostras de folhas de plantas adultas coletadas em seus habitats nativos de restinga (Praia Seca no Rio de Janeiro). Neste ambiente as pitangueiras estão distribuídas em "*patches*" de vários tamanhos separadas por areia (Pimentel et al. 2007) e sujeitas a diferentes tipos de estresse presentes em muitas restingas, como previamente mencionado. Das 11 famílias de miRNAs avaliados, sete delas mostraram um padrão diferente de expressão dependendo do momento do dia. Ao meio-dia, só o miR170 mostrou um padrão de regulação positiva, enquanto os demais foram regulados negativamente. Também foi avaliada a expressão desses miRNAs, que potencialmente estariam ajudando a pitanga na tolerância ao estresse, em condições laboratoriais nas quais as plantas estariam sendo submetidas a somente um tipo de estresse. Portanto, a segunda condição foi submeter as plantas a condições controladas de seca ou défice hídrico e osmótico induzido por PEG e coletar as folhas para a posterior análise por RT-qPCR. Nas análises de expressão, observou-se que cinco miRNAs mostraram expressão diferencial, dos quais quatro deles (miR166, miR170, miR172 e miR396) também foram diferencialmente expressos em amostras coletadas diretamente na restinga. Todos esses miRNAs foram previamente reportados em outras espécies, como por exemplo o miR166 e miR396 que mostraram uma regulação negativa em *O. sativa* e num cultivar de soja tolerante sob condições de seca (Zhou et al. 2010; Kulcheski et al. 2011). O miR170 também foi descrito como um miRNA induzido por seca, em estudo de genoma completo de *O. sativa* (Zahid et al. 2016). Já o

miR172 foi descrito em *A. thaliana* como aumentando a tolerância ao estresse hídrico e tolerância a salinidade (Li et al. 2016).

Os miRNAs também estão envolvidos em processos metabólicos e de desenvolvimento das plantas (Li and Zhang 2016). Portanto, o conhecimento do conjunto completo de miRNAs de uma espécie é importante para entender os mecanismos complexos de regulação que acontecem nela. Assim, os novos miRNAs foram testados em diferentes tecidos de pitanga. Dos 20 miRNAs testados, 14 mostraram padrões diferenciais de expressão, sendo eun-miR10216-5p, eun-nMIR002-5p, eun-nMIR005-3p, eun-nMIR006-5p e eun-nMIR012-3p candidatos a serem específicos de pétala; eun-nMIR011-5p e eun-nMIR013-5p específicos de raiz e eun-miR10218-5p e eun-nMIR001-5p como específicos de folhas. O padrão tecido-especifico dos miRNAs tem sido amplamente descrito para sncRNAs (Pappas et al. 2015). Os alvos preditos destes miRNAs foram categorizados com sucesso em processos biológicos e funções moleculares por análises de Ontologia usando o Blast2Go. Entre os mRNA alvos, à exceção dos que codificam fatores de transcrição, foram encontrados genes com as mais variadas funções como metiltransferases, cinases ou hidrolases. Estas observações são similares a outros trabalhos que identificam alvos pouco conservados de novos miRNAs (Jeong et al. 2011; Hwang et al. 2013). Alguns deles foram avaliados nos diferentes tecidos e mostraram o padrão de expressão oposto ao dos miRNAs indicando uma inibição pós-transcricional destes genes.

Os resultados obtidos neste trabalho sugerem que os miRNAs tem relação com o desenvolvimento de *E. uniflora*, possivelmente controlando a formação das folhas, mudando a arquitetura de folhas em pétalas, atuando em resposta ao estresse ou participando de outras vias de sinalização. Porém, para obter-se uma melhor informação sobre a interação do miRNA com seu alvo, o sequenciamento do degradoma de diferentes tecidos poderia ser uma alternativa para completar nossa compreensão sobre o papel dos miRNAs e seus alvos.

No segundo artigo, decidimos expandir a pesquisa para os fragmentos derivados de RNA transportador (tRFs) com a finalidade de tentar completar o cenário de regulação por sncRNAs nas pitangas. Os tRFs tem sido descritos na resposta ao estresse biótico (Asha and Soniya 2016) e abiótico (Hsieh et al.

2009; Hackenberg et al. 2013; Loss-Morais et al. 2013; Alves et al. 2017). Com essa finalidade, os tRNAs da pitanga (410 tRNAs) foram anotados usando a mesma versão do genoma usado nos miRNAs e decidimos nos focar nos 43 tRNAs ortólogos com *E. grandis*. Assim foi possível identificar 469 tRFs em *E. uniflora* contra os 273 tRFs em *E. grandis*. Os tRFs foram caracterizados segundo a abundância, observando-se uma predominância de 5'tRFs de 18 nt para *E. uniflora* e de 24 nt para *E. grandis*, corroborando a predominância de mapeamento de sncRNAs na extremidade 5' do tRNA e a variabilidade no comprimento deles que parece ser dependente da espécie (Kumar et al. 2014; Alves et al. 2017), também foi caracterizado a preferência de nucleotídeos no sítio de clivagem e observado um padrão de abundância de timinas nos 5'tRFs ausente nos 3'tRFs. Todas essas características indicam que essas sequências não são simples produtos de degradação e sim sncRNAs específicos e fortemente regulados.

Tendo em mente que a seca e a salinidade são os dois tipos de estresses abióticos que mais influenciam a produção das culturas e a qualidade das sementes; e além disso porque pitanga cresce em condições ambientais parecidas, decidimos avaliar os tRFs em seca e salinidade. A seca foi induzida pelo PEG800 por períodos longos de estresses de 48 horas e uma semana, enquanto que o cloreto de sódio (NaCl) na concentração de 200 mM serviu para simular condições de estresse salino por períodos mais curtos de quatro horas e 48 horas. Assim foram avaliados 11 tRFs conservados em *E. uniflora* e em *E. grandis* mas que se diferenciavam na abundância. Os resultados do RT-qPCR mostraram que seis deles variam significativamente em algum momento tanto para salinidade como para seca. Identificamos alguns tRFs específicos de um tipo estresse como o 5'tRF AlaTGC de salinidade e o 5'tRF SerCGA de seca.

O passo seguinte foi avaliar os alvos potenciais destes tRFs. Para isso, foram escolhidos aqueles tRFs com padrão similar de expressão em ambos estresses (5'tRF ArgTCG e 5'tRF GlyTCC) e com padrão oposto como o 5'tRF GlyCCC. Os alvos foram previamente identificados *in silico* como fatores de transcrição, envolvidos em condições de crescimento e em estresse. Depois da análise do padrão de expressão em ambos estresses, mRNAs indicados como alvos mostraram inibição na expressão deles devido ao processo de inibição

pós-transcricional. Genes como F-box/LRR repeat domain, DEAD-box (fatores de transcrição), transportadores de magnésio, helicases, metaloproteases ou proteínas envolvidas em modificações epigenéticas como demetilases e methiltransferases foram preditos. Todos eles foram previamente identificados em outras espécies sob estresse abiótico embora não tenham exatamente o mesmo padrão de expressão (Moreno et al. 2005; Gupta et al. 2015; Shchennikova et al. 2016). Isso pode ser explicado porque a reposta da planta ao estresse muitas vezes depende da espécie, o tecido onde acontece o estresse e a duração dele, tornando esse processo muito mais complexo (Chaves et al. 2009). Embora alguns processos celulares e metabólicos observados durante a seca e a salinidade são semelhantes, existem muitos genes e rotas metabólicas que discriminam essas duas condições de estresse (Bartels and Sunkar 2005). Isso estaria explicando porque não foram achadas diferenças significativas nos alvos do 5'tRF GlyCCC em salinidade.

Esses resultados, tanto os obtidos com os miRNAs como com os tRFs, podem servir para futuros trabalhos que integrem proteômica e metabolômica com a finalidade de obter uma melhor compreensão das sofisticadas e finamente reguladas redes moleculares que envolvem a tolerância a seca e salinidade presente em *E. uniflora*. Nesse sentido são necessários mais estudos para obter um maior conhecimento sobre os genes e proteínas que são seletivamente regulados na resposta ao estresse. Porém, a participação de sncRNAs como tRFs e miRNA foi evidenciada.

4. REFERÊNCIAS

Almeida DJ De, Faria MV and Silva PR Da (2012) Biologia experimental em Pitangueira: uma revisão de cinco décadas de publicações científicas / Experimental biology in pitangueira: a review of five decades of scientific publications. Rev Ambiência 8:159–175. doi: 10.5777/ambiencia.2012.01.02rb

Alves CS, Vicentini R, Duarte GT, Pinoti VF, Vincentz M and Nogueira FTS (2017) Genome-wide identification and characterization of tRNA-derived RNA fragments in land plants. Plant Mol Biol 93:35–48. doi: 10.1007/s11103-016-0545-9

Arenas-Huertero C, Pérez B, Rabanal F, Blanco-Melo D, De La Rosa C, Estrada-Navarrete G, Sanchez F, Covarrubias AA and Reyes JL (2009) Conserved and novel miRNAs in the legume Phaseolus vulgaris in response to stress. Plant Mol Biol 70:385–401. doi: 10.1007/s11103-009-9480-3

Asha S and Soniya E V. (2016) Transfer RNA Derived Small RNAs Targeting Defense Responsive Genes Are Induced during Phytophthora capsici Infection in Black Pepper (Piper nigrum L.). Front Plant Sci. doi: 10.3389/fpls.2016.00767

Åsman A, Vetukuri RR, Avrova AO, Jahan SN, Whisson SC, Fogelqvist J, Corcoran P, Avrova AO, Whisson SC and Dixelius C (2014) Fragmentation of tRNA in *Phytophthora infestans* asexual life cycle stages and during host plant infection. BMC Microbiol 14:308. doi: 10.1186/s12866-014-0308-1

Axtell MJ and Meyers BC (2018) Revisiting criteria for plant miRNA annotation in the era of big data. Plant Cell tpc.00851.2017. doi: 10.1105/tpc.17.00851

Barrera-Figueroa BE, Gao L, Diop NN, Wu Z, Ehlers JD, Roberts PA, Close TJ, Zhu JK and Liu R (2011) Identification and comparative analysis of drought-associated microRNAs in two cowpea genotypes. BMC Plant Biol. doi: 10.1186/1471-2229-11-127

Bartel DP (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell 116:281–297. doi: 10.1016/S0092-8674(04)00045-5

Bartels D and Sunkar R (2005) Drought and salt tolerance in plants. CRC Crit Rev Plant Sci 24:23–58. doi: 10.1080/07352680590910410

Bologna NG and Voinnet O (2014) The Diversity, Biogenesis, and Activities of Endogenous Silencing Small RNAs in *Arabidopsis*. Annu Rev Plant Biol 65:473–503. doi: 10.1146/annurev-arplant-050213-035728

Bonnet E, Wuyts J, Rouze P and Van de Peer Y (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. Proc Natl Acad Sci 101:11511–11516. doi: 10.1073/pnas.0404025101

Bühler M, Spies N, Bartel DP and Moazed D (2008) TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway. Nat Struct Mol Biol 15:1015–1023. doi: 10.1038/nsmb.1481

Cai P, Piao X, Hao L, Liu S, Hou N, Wang H and Chen Q (2013) A Deep Analysis of the Small Non-Coding RNA Population in Schistosoma japonicum Eggs. PLoS One. doi: 10.1371/journal.pone.0064003

Chaves MM, Flexas J and Pinheiro C (2009) Photosynthesis under drought and salt stress: Regulation mechanisms from whole plant to cell. Ann Bot 103:551–560. doi: 10.1093/aob/mcn125

Chen C-J, liu Q, Zhang Y-C, Qu L-H, Chen Y-Q and Gautheret D (2011) Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embryogenic callus. RNA Biol 8:538–547. doi: 10.4161/rna.8.3.15199

Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR et al. (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res. doi: 10.1093/nar/gni178

Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JWS, Green PJ, Barton GJ and Hutvagner G (2009) Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA 15:2147–2160. doi: 10.1261/rna.1738409

Couvillion MT, Sachidanandam R and Collins K (2010) A growth-essential Tetrahymena Piwi protein carries tRNA fragment cargo. Genes Dev 24:2742–2747. doi: 10.1101/gad.1996210

Desvignes T, Batzel P, Berezikov E, Eilbeck K, Eppig JT, McAndrews MS, Singer A and Postlethwait JH (2015) MiRNA Nomenclature: A View Incorporating Genetic Origins, Biosynthetic Pathways, and Sequence Variants. Trends Genet 31:613–626. doi: 10.1016/j.tig.2015.09.002

Friedländer MR, MacKowiak SD, Li N, Chen W and Rajewsky N (2012) MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40:37–52. doi: 10.1093/nar/gkr688

Fu C, Sunkar R, Zhou C, Shen H, Zhang JY, Matts J, Wolf J, Mann DGJ, Stewart CN, Tang Y et al. (2012) Overexpression of miR156 in switchgrass (Panicum virgatum L.) results in various morphological alterations and leads to improved biomass production. Plant Biotechnol J 10:443–452. doi: 10.1111/j.1467-7652.2011.00677.x

Gebetsberger J, Zywicki M, Künzi A and Polacek N (2012) TRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in Haloferax volcanii. Archaea. doi: 10.1155/2012/260909

Goswami S, Kumar RR and Rai RD (2014) Heat-responsive microRNAs regulate the transcription factors and heat shock proteins in modulating thermo-stability of starch biosynthesis enzymes in wheat (Triticum aestivum L.) under the heat stress. Aust J Crop Sci 8:697–705.

Govaerts R, Sobral M, Ashton P, Barrie F, Holst BK, Landrum LL, Matsumoto K, Mazine FF, Lughadha EN, Proenca C et al. (2015) World Checklist of Myrtaceae. In: R. Bot. Gard. Kew. http://apps.kew.org/wcsp/synonomy.do;jsessionid=C2E72FE08A14CAE1130BB94B19007663?name_id=80144.

Griffiths-Jones S, Saini HK, Van Dongen S and Enright AJ (2008) miRBase: Tools for microRNA genomics. Nucleic Acids Res. doi: 10.1093/nar/gkm952

Gupta OP, Sharma P, Gupta RK and Sharma I (2014) MicroRNA mediated regulation of metal toxicity in plants: Present status and future perspectives. Plant Mol Biol 84:1–18. doi: 10.1007/s11103-013-0120-6

Gupta S, Garg V, Kant C and Bhatia S (2015) Genome-wide survey and expression analysis of F-box genes in chickpea. BMC Genomics 16:67. doi:

10.1186/s12864-015-1293-y

Guzman F, Almerão MP, Körbes AP, Loss-Morais G and Margis R (2012) Identification of MicroRNAs from Eugenia uniflora by High-Throughput Sequencing and Bioinformatics Analysis. PLoS One. doi: 10.1371/journal.pone.0049811

Hackenberg M, Huang PJ, Huang CY, Shi BJ, Gustafson P and Langridge P (2013) A Comprehensive expression profile of micrornas and other classes of non-coding small RNAs in barley under phosphorous-deficient and-sufficient conditions. DNA Res 20:109–125. doi: 10.1093/dnares/dss037

Haussecker D, Huang Y, Lau A, Parameswaran P, Fire AZ and Kay MA (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA 16:673–695. doi: 10.1261/rna.2000810

Heo JB, Lee Y-S and Sung S (2013) Epigenetic regulation by long noncoding RNAs in plants. Chromosom Res 21:685–693. doi: 10.1007/s10577-013-9392-6

Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–3431. doi: 10.1093/nar/gkg599

Hong Y and Jackson S (2015) Floral induction and flower formation-the role and potential applications of miRNAs. Plant Biotechnol J 13:282–292. doi: 10.1111/pbi.12340

Hsieh L-C, Lin S-I, Shih AC-C, Chen J-W, Lin W-Y, Tseng C-Y, Li W-H and Chiou T-J (2009) Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiol 151:2120–2132. doi: 10.1104/pp.109.147280

Hwang DG, Park JH, Lim JY, Kim D, Choi Y, Kim S, Reeves G, Yeom SI, Lee JS, Park M et al. (2013) The Hot Pepper (Capsicum annuum) MicroRNA Transcriptome Reveals Novel and Conserved Targets: A Foundation for Understanding MicroRNA Functional Roles in Hot Pepper. PLoS One. doi: 10.1371/journal.pone.0064238

Itaya A, Bundschuh R, Archual AJ, Joung J-G, Fei Z, Dai X, Zhao PX, Tang Y, Nelson RS and Ding B (2008) Small RNAs in tomato fruit and leaf development.

Biochim Biophys Acta 1779:99–107. doi: 10.1016/j.bbagrm.2007.09.003

Ivanov P, Emara MM, Villen J, Gygi SP and Anderson P (2011) Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation. Mol Cell 43:613–623. doi: 10.1016/j.molcel.2011.06.022

Jeong D-HD, Park S, Zhai J, Gurazada SGR, De Paoli E, Meyers BC and Green PJ (2011) Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. Plant Cell 23:4185–207. doi: 10.1105/tpc.111.089045

Jian X, Zhang L, Li G, Zhang L, Wang X, Cao X, Fang X and Chen F (2010) Identification of novel stress-regulated microRNAs from Oryza sativa L. Genomics 95:47–55. doi: 10.1016/j.ygeno.2009.08.017

José Ripoll J, Bailey LJ, Mai Q-A, Wu SL, Hon CT, Chapman EJ, Ditta GS, Estelle M and Yanofsky MF (2015) microRNA regulation of fruit growth. Nat Plants 1:15036. doi: 10.1038/nplants.2015.36

Kamthan A, Chaudhuri A, Kamthan M and Datta A (2015) Small RNAs in plants: recent development and application for crop improvement. Front Plant Sci. doi: 10.3389/fpls.2015.00208

Kantar M, Lucas SJ and Budak H (2011) miRNA expression patterns of Triticum dicoccoides in response to shock drought stress. Planta 233:471–484. doi: 10.1007/s00425-010-1309-4

Kawaji H, Nakamura M, Takahashi Y, Sandelin A, Katayama S, Fukuda S, Daub CO, Kai C, Kawai J, Yasuda J et al. (2008) Hidden layers of human small RNAs. BMC Genomics. doi: 10.1186/1471-2164-9-157

Keam S and Hutvagner G (2015) tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. Life 5:1638–1651. doi: 10.3390/life5041638

Khan GA, Declerck M, Sorin C, Hartmann C, Crespi M and Lelandais-Brière C (2011) MicroRNAs as regulators of root development and architecture. Plant Mol Biol 77:47–58. doi: 10.1007/s11103-011-9793-x

Kulcheski FR, de Oliveira LF, Molina LG, Almerão MP, Rodrigues FA, Marcolino

J, Barbosa JF, Stolf-Moreira R, Nepomuceno AL, Marcelino-Guimarães FC et al. (2011) Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics 12:307. doi: 10.1186/1471-2164-12-307

Kumar P, Anaya J, Mudunuri SB and Dutta A (2014) Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Med. doi: 10.1186/s12915-014-0078-0

Kumar P, Mudunuri SB, Anaya J and Dutta A (2015) tRFdb: A database for transfer RNA fragments. Nucleic Acids Res 43:D141–D145. doi: 10.1093/nar/gku1138

Lee RC, Feinbaum RL and Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854. doi: 10.1016/0092-8674(93)90529-Y

Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH and Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–4060. doi: 10.1038/sj.emboj.7600385

Lee YS, Shibata Y, Malhotra A and Dutta A (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev 23:2639–2649. doi: 10.1101/gad.1837609

Lei J and Sun Y (2014) MiR-PREFeR: An accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. Bioinformatics 30:2837–2839. doi: 10.1093/bioinformatics/btu380

Levy A, Szwerdszarf D, Abu-Abied M, Mordehaev I, Yaniv Y, Riov J, Arazi T and Sadot E (2014) Profiling microRNAs in Eucalyptus grandis reveals no mutual relationship between alterations in miR156 and miR172 expression and adventitious root induction during development. BMC Genomics. doi: 10.1186/1471-2164-15-524

Li C and Zhang B (2016) MicroRNAs in Control of Plant Development. J Cell Physiol 231:303–313. doi: 10.1002/jcp.25125

Li W, Wang T, Zhang Y and Li Y (2016) Overexpression of soybean miR172c

confers tolerance to water deficit and salt stress, but increases ABA sensitivity in transgenic Arabidopsis thaliana. J Exp Bot 67:175–194. doi: 10.1093/jxb/erv450

Liang G, He H and Yu D (2012) Identification of Nitrogen Starvation-Responsive MicroRNAs in Arabidopsis thaliana. PLoS One. doi: 10.1371/journal.pone.0048951

Liao J-Y, Guo Y-H, Zheng L-L, Li Y, Xu W-L, Zhang Y-C, Zhou H, Lun Z-R, Ayala FJ and Qu L-H (2014) Both endo-siRNAs and tRNA-derived small RNAs are involved in the differentiation of primitive eukaryote *Giardia lamblia*. Proc Natl Acad Sci 111:14159–14164. doi: 10.1073/pnas.1414394111

Liao JY, Ma LM, Guo YH, Zhang YC, Zhou H, Shao P, Chen YQ and Qu LH (2010) Deep sequencing of human nuclear and cytoplasmic small RNAS reveals an unexpectedly complex subcellular distribution of mirnas and tRNA 3′ trailers. PLoS One. doi: 10.1371/journal.pone.0010563

Lim TK (2012) Edible Medicinal And Non-Medicinal Plants. Edible Med Non-Medicinal Plants 2:867–878. doi: 10.1007/978-94-007-1764-0

Liu H-H, Tian X, Li Y-J, Wu C-A and Zheng C-C (2008) Microarray-based analysis of stress-regulated microRNAs in Arabidopsis thaliana. RNA 14:836–843. doi: 10.1261/rna.895308

Liu M, Yu H, Zhao G, Huang Q, Lu Y and Ouyang B (2017) Profiling of drought-responsive microRNA and mRNA in tomato using high-throughput sequencing. BMC Genomics. doi: 10.1186/s12864-017-3869-1

Loss-Morais G, Ferreira DCR, Margis R, Alves-Ferreira M and Corrêa RL (2014) Identification of novel and conserved MicroRNAs in coffea canephora and coffea arabica. Genet Mol Biol 37:671–682. doi: 10.1590/S1415-47572014005000020

Loss-Morais G, Waterhouse PM and Margis R (2013) Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. Biol Direct 8:6. doi: 10.1186/1745-6150-8-6

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC and Green PJ (2005) Genetics: Elucidation of the small RNA component of the transcriptome. Science (80-) 309:1567–1569. doi: 10.1126/science.1114112

Lucas EJ and Bünger MO (2015) Myrtaceae in the Atlantic forest: their role as a “model” group. Biodivers Conserv 24:2165–2180. doi: 10.1007/s10531-015-0992-7

Ma X, Tang Z, Qin J and Meng Y (2015) The use of high-throughput sequencing methods for plant microRNA research. RNA Biol 12:709–719. doi: 10.1080/15476286.2015.1053686

Margis R, Felix D, Caldas JF, Salgueiro F, De Araujo DSD, Breyne P, Van Montagu M, De Oliveira D and Margis-Pinheiro M (2002) Genetic differentiation among three neighboring Brazil-cherry (Eugenia uniflora L.) populations within the Brazilian Atlantic rain forest. Biodivers Conserv 11:149–163. doi: 10.1023/A:1014028026273

Martinez G, Choudury SG and Slotkin RK (2017) TRNA-derived small RNAs target transposable element transcripts. Nucleic Acids Res 45:5142–5152. doi: 10.1093/nar/gkx103

Mazine FF, Souza VC, Sobral M, Forest F and Lucas E (2014) A preliminary phylogenetic analysis of Eugenia (Myrtaceae: Myrteae), with a focus on Neotropical species. Kew Bull 69:1–14. doi: 10.1007/s12225-014-9497-x

Mcvaugh R (1968) The Genera of American Myrtaceae: An Interim Report. Source: Taxon 17:354–418. doi: 10.2307/1217393

Meng Y, Shao C and Chen M (2011) Toward microRNA-mediated gene regulatory networks in plants. Brief Bioinform 12:645–659. doi: 10.1093/bib/bbq091

Moreno JI, Martín R and Castresana C (2005) Arabidopsis SHMT1, a serine hydroxymethyltransferase that functions in the photorespiratory pathway influences resistance to biotic and abiotic stress. Plant J 41:451–463. doi: 10.1111/j.1365-313X.2004.02311.x

Myers N, Myers N, Mittermeier R a, Mittermeier R a, Fonseca G a B, Fonseca G a B, Kent J and Kent J (2000) Biodiversity hotspots for conservation priorities. Nature 403:853–8. doi: 10.1038/35002501

Nozawa M, Miura S and Nei M (2012) Origins and evolution of microRNA genes

in plant species. Genome Biol Evol 4:230–239. doi: 10.1093/gbe/evs002

Ogawa T, Tomita K, Ueda T, Watanabe K, Uozumi T and Masaki H (1999) A cytotoxic ribonuclease targeting specific transfer RNA anticodons. Science (80-) 283:2097–2100. doi: 10.1126/science.283.5410.2097

Oliveira-Filho AT and Fontes MAL (2000) Patterns of Floristic Differentiation among Atlantic Forests in Southeastern Brazil and the Influence of Climate1. Biotropica 32:793–810. doi: 10.1111/j.1744-7429.2000.tb00619.x

Pappas M de CR, Pappas GJ and Grattapaglia D (2015) Genome-wide discovery and validation of Eucalyptus small RNAs reveals variable patterns of conservation and diversity across species of Myrtaceae. BMC Genomics. doi: 10.1186/s12864-015-2322-6

Pimentel MCP, Barros MJ, Cirne P, Mattos E a. De, Oliveira RC, Pereira MC a., Scarano FR, Zaluar HLT and Araujo DSD (2007) Spatial variation in the structure and floristic composition of “restinga” vegetation in southeastern Brazil. Rev Bras Botânica 30:543–551. doi: 10.1590/S0100-84042007000300018

Ribeiro MC, Metzger JP, Martensen AC, Ponzoni FJ and Hirota MM (2009) The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. Biol Conserv 142:1141–1153. doi: 10.1016/j.biocon.2009.02.021

Roesch LFW, Vieira FCB, Pereira VA, Schünemann AL, Teixeira IF, Senna AJT and Stefenon VM (2009) The Brazilian Pampa: A fragile biome. Diversity 1:182–198. doi: 10.3390/d1020182

Rubio-Somoza I and Weigel D (2011) MicroRNA networks and developmental plasticity in plants. Trends Plant Sci 16:258–264. doi: 10.1016/j.tplants.2011.03.001

Salgueiro F, Felix D, Caldas JF, Margis-Pinheiro M and Margis R (2004) Even population differentiation for maternal and biparental gene markers in Eugenia uniflora, a widely distributed species from the Brazilian coastal Atlantic rain forest. Divers Distrib 10:201–210. doi: 10.1111/j.1366-9516.2004.00078.x

Scarano FR (2002) Structure, function and floristic relationships of plant

communities in stressful habitats marginal to the Brazilian Atlantic rainforest. Ann Bot 90:517–524. doi: 10.1093/aob/mcf189

Scarano FR, Duarte HM, Ribeiro KT, Rodrigues PJFP, Barcellos EMB, Franco AC, Brulfert J, Deléens E and Lüttge U (2001) Four sites with contrasting environmental stress in southeastern Brazil: Relations of species, life form diversity, and geographic distribution to ecophysiological parameters. Bot J Linn Soc 136:345–364. doi: 10.1006/bojl.2000.0435

Shchennikova A V., Beletsky A V., Shulga OA, Mazur AM, Prokhortchouk EB, Kochieva EZ, Ravin N V. and Skryabin KG (2016) Deep-sequence profiling of miRNAs and their target prediction in Monotropa hypopitys. Plant Mol Biol 91:441–458. doi: 10.1007/s11103-016-0478-3

Shen EM, Singh SK, Ghosh JS, Patra B, Paul P, Yuan L and Pattanaik S (2017) The miRNAome of Catharanthus roseus: Identification, expression analysis, and potential roles of microRNAs in regulation of terpenoid indole alkaloid biosynthesis. Sci Rep. doi: 10.1038/srep43027

Soares AR and Santos M (2017) Discovery and function of transfer RNA-derived fragments and their role in disease. Wiley Interdiscip Rev RNA. doi: 10.1002/wrna.1423

Sobala A and Hutvagner G (2013) Small RNAs derived from the 5’ end of tRNA can inhibit protein translation in human cells. RNA Biol 10:553–63. doi: 10.4161/rna.24285

Spada PDS, de Souza GGN, Bortolini GV, Henriques JAP and Salvador M (2008) Antioxidant, mutagenic, and antimutagenic activity of frozen fruits. J Med Food 11:144–151. doi: 10.1089/jmf.2007.598

Stehmann JR, Forzza RC, Salino A, Sobral M, Costa DP and Kamino LHY (2009) Plantas da floresta Atlântica.

Telonis AG, Loher P, Honda S, Jing Y, Palazzo J, Kirino Y and Rigoutsos I (2015) Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. Oncotarget 6:24797–24822. doi: 10.18632/oncotarget.4695

Tomita K, Ogawa T, Uozumi T, Watanabe K and Masaki H (2000) A cytotoxic ribonuclease which specifically cleaves four isoaccepting arginine tRNAs at their anticodon loops. Proc Natl Acad Sci U S A 97:8278–8283. doi: 10.1073/pnas.140213797

Toscano S, Farieri E, Ferrante A and Romano D (2016) Physiological and Biochemical Responses in Two Ornamental Shrubs to Drought Stress. Front Plant Sci. doi: 10.3389/fpls.2016.00645

Trindade I, Capitão C, Dalmay T, Fevereiro MP and dos Santos DM (2010) miR398 and miR408 are up-regulated in response to water deficit in Medicago truncatula. Planta 231:705–716. doi: 10.1007/s00425-009-1078-0

Turchetto-Zolet AC, Salgueiro F, Turchetto C, Cruz F, Veto NM, Barros MJF, Segatto ALA, Freitas LB and Margis R (2016) Phylogeography and ecological niche modelling in Eugenia uniflora (Myrtaceae) suggest distinct vegetational responses to climate change between the southern and the northern Atlantic Forest. Bot J Linn Soc 182:670–688. doi: 10.1111/boj.12473

Wang B, Sun YF, Song N, Wang XJ, Feng H, Huang LL and Kang ZS (2013) Identification of UV-B-induced microRNAs in wheat. Genet Mol Res 12:4213–4221. doi: 10.4238/2013.October.7.7

Wang JW, Park MY, Wang LJ, Koo Y, Chen XY, Weigel D and Poethig RS (2011a) MiRNA control of vegetative phase change in trees. PLoS Genet. doi: 10.1371/journal.pgen.1002012

Wang L, Yu X, Wang H, Lu Y-Z, de Ruiter M, Prins M and He Y-K (2011b) A novel class of heat-responsive small RNAs derived from the chloroplast genome of Chinese cabbage (Brassica rapa). BMC Genomics 12:289. doi: 10.1186/1471-2164-12-289

Wang Q, Li T, Xu K, Zhang W, Wang X, Quan J, Jin W, Zhang M, Fan G, Wang M-B et al. (2016a) The tRNA-Derived Small RNAs Regulate Gene Expression through Triggering Sequence-Specific Degradation of Target Transcripts in the Oomycete Pathogen Phytophthora sojae. Front Plant Sci 7:1938. doi: 10.3389/fpls.2016.01938

Wang Y, Li H, Sun Q and Yao Y (2016b) Characterization of small RNAs derived

from tRNAs, rRNAs and snoRNAs and their response to heat stress in wheat seedlings. PLoS One. doi: 10.1371/journal.pone.0150933

Wen M, Xie M, He L, Wang Y, Shi S and Tang T (2016) Expression Variations of miRNAs and mRNAs in Rice (Oryza sativa). Genome Biol Evol 8:3529–3544. doi: 10.1093/gbe/evw252

Wightman B, Ha I and Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 75:855–862. doi: 10.1016/0092-8674(93)90530-4

Wilson PG, O'Brien MM, Gadek PA and Quinn CJ (2001) Myrtaceae revisited: a reassessment of infrafamilial groups. Am J Bot 88:2013–25.

Wilson PG, O'Brien MM, Heslewood MM and Quinn CJ (2005) Relationships within Myrtaceae sensu lato based on a matK phylogeny. Plant Syst Evol 251:3–19. doi: 10.1007/s00606-004-0162-y

Xie F, Wang Q, Sun R and Zhang B (2015) Deep sequencing reveals important roles of microRNAs in response to drought and salinity stress in cotton. J Exp Bot 66:789–804. doi: 10.1093/jxb/eru437

Yang X and Li L (2012) Analyzing the microRNA Transcriptome in Plants Using Deep Sequencing Data. Biology (Basel) 1:297–310. doi: 10.3390/biology1020297

Yang X, Zhang H and Li L (2011) Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data. Genomics 98:40–46. doi: 10.1016/j.ygeno.2011.03.011

Zahid KR, Ali F, Shah F, Younas M, Shah T, Shahwar D, Hassan W, Ahmad Z, Qi C, Lu Y et al. (2016) Response and Tolerance Mechanism of Cotton Gossypium hirsutum L. to Elevated Temperature Stress: A Review. Front Plant Sci. doi: 10.3389/fpls.2016.00937

Zhang B (2015) MicroRNA: A new target for improving plant tolerance to abiotic stress. J Exp Bot 66:1749–1761. doi: 10.1093/jxb/erv013

Zhang B, Pan X, Cobb GP and Anderson TA (2006) Plant microRNA: A small regulatory molecule with big impact. Dev Biol 289:3–16. doi:

10.1016/j.ydbio.2005.10.036

Zhang B and Wang Q (2015) MicroRNA-based biotechnology for plant improvement. J Cell Physiol 230:1–15. doi: 10.1002/jcp.24685

Zhang BH, Pan XP, Wang QL, Cobb GP and Anderson TA (2005) Identification and characterization of new plant microRNAs using EST analysis. Cell Res 15:336–360. doi: 10.1038/sj.cr.7290302

Zhang H, Wan Q, Ye W, Lv Y, Wu H and Zhang T (2013) Genome-Wide Analysis of Small RNA and Novel MicroRNA Discovery during Fiber and Seed Initial Development in Gossypium hirsutum. L. PLoS One. doi: 10.1371/journal.pone.0069743

Zhang S, Sun L and Kragler F (2009) The Phloem-Delivered RNA Pool Contains Small Noncoding RNAs and Interferes with Translation. PLANT Physiol 150:378–387. doi: 10.1104/pp.108.134767

Zhou L, Liu Y, Liu Z, Kong D, Duan M and Luo L (2010) Genome-wide identification and analysis of drought-responsive microRNAs in Oryza sativa. J Exp Bot 61:4157–4168. doi: 10.1093/jxb/erq237

5. ANEXOS

Outras produções cientificas relacionadas no período

Plant Syst Evol (2017) 303:1199–1212
DOI 10.1007/s00606-017-1431-x

ORIGINAL ARTICLE

The chloroplast genome sequence from *Eugenia uniflora*, a Myrtaceae from Neotropics

Maria Eguiluz[1] · Nureyev F. Rodrigues[1] · Frank Guzman[2] · Priscila Yuyama[1] · Rogerio Margis[1,2,3]

Received: 28 September 2016 / Accepted: 28 May 2017 / Published online: 19 June 2017

Abstract *Eugenia uniflora* is a plant native to tropical America that holds great ecological and economic importance. The complete chloroplast (cp) genome sequence of *Eugenia uniflora*, a member of the Neotropical Myrtaceae family, is reported here. The genome is 158,445 bp in length and exhibits a typical quadripartite structure of the large (LSC, 87,459 bp) and small (SSC, 18,318 bp) single-copy regions, separated by a pair of inverted repeats (IRs, 26,334 bp). It contains 111 unique genes, including 77 protein-coding genes, 30 tRNAs and 4 rRNAs. The genome structure, gene order, GC content and codon usage are similar to the typical angiosperm cp genomes. Comparison of the entire cp genomes of *E. uniflora* L. and three other Myrtaceae revealed an expansion of 43 bp in the intergenic spacer located between the IRA/large single-copy (LSC) border and the first gene of LSC region. Simple sequence repeat (SSR) analysis revealed that most SSRs are AT rich, which contribute to the overall AT richness of the cp genome. Additionally, fewer SSRs are distributed in the protein-coding sequences compared to the noncoding regions. Phylogenetic analysis among 58 species based on 57 cp genes demonstrated a closer relationship between *E. uniflora* L. and *Syzygium cumini* (L). Skeels compared to the Eucalyptus clade in the Myrtaceae family. The complete cp genome sequence of *E. uniflora* reported here has importance for population genetics, as well as phylogenetic and evolutionary studies in this species and other Myrtaceae species from Neotropical regions.

Keywords cpDNA · Fruit tree · Genome sequencing · NGS · Pitanga · Plant evolution

Handling editor: Marcus Koch.

Electronic supplementary material The online version of this article (doi:10.1007/s00606-017-1431-x) contains supplementary material, which is available to authorized users.

✉ Rogerio Margis
rogerio.margis@ufrgs.br

[1] PPGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

[2] PPGBCM, Centro de Biotecnologia, sala 213, prédio 43431, Universidade Federal do Rio Grande do Sul - UFRGS, PO Box 15005, Porto Alegre, RS CEP 91501-970, Brazil

[3] Departamento de Biofísica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

Introduction

Chloroplasts are multifunctional organelles, which possess their own genetic material and are believed to have originated from ancient endosymbiotic cyanobacteria (Ravi et al. 2008). The chloroplast (cp) genome in angiosperms usually varies between 115 and 165 kb in size and maintains highly conserved organization in most land plants. The lack of recombination, low rates of nucleotide substitutions (Wolfe et al. 1987) and primarily uniparental inheritance make plant cpDNA a valuable genetic source for phylogenetic relationship studies (Bayly et al. 2013). Sequence data from the plastid genome have transformed plant systematics and contributed greatly to unravel deep-level evolutionary relationships of taxonomically unresolved plant taxa (Jansen et al. 2007; Moore et al. 2010; Ruhfel et al. 2014).

The Myrtaceae (Myrtle, Eucalyptus, clove or guava family) is the eighth largest flowering plant family, and it is dominant among several vegetation types in South America through a variety of ecotypes (Pennington et al. 2009). *Eugenia* is the largest genus of Neotropical Myrtaceae

family, encompassing about 5600 species, two-thirds of which are present in Brazilian ecosystems (Govaerts et al. 2015). *Eugenia* can be distinguished from the other genera of tribe Myrteae DC. by the generally 4-merous flowers, which have free calyx lobes that are separate in the flower bud, a non-tubular hypanthium that usually not extend beyond the tip of the bilocular multiovulate ovary, and finally by their embryo with cotyledons fused in a solid homogeneous mass (Mazine et al. 2014).

Eugenia uniflora L. is a fruit tree native to South America that serves as a good model for ecological studies because it grows in several different vegetation types, including forests, restingas, and arid and semiarid environments in the Brazilian northeast. This species is very versatile in terms of adaptability and plays a fundamental role in the maintenance of the shrubby coastal vegetation. Ecologically, it is an important food source for a variety of birds and mammals, and it can survive in disturbed sites within restinga habitats, especially near the beach (Almeida et al. 2012). Besides its ecologic importance, *E. uniflora* L. produces edible cherry-like fruits characterized by a low lipid and caloric content and by high amounts of polyphenols, carotenoids, and other antioxidant compounds (Spada et al. 2008) being traditionally used in folk medicine as antipyretic, stomachic, hypoglycemic, and to lower blood pressure (Lim 2012).

Despite the importance of the family, the phylogenetic relationships and delimitation of some genera are still debatable, especially in the fleshy fruit members. Many studies have provided insights into Myrtaceae phylogeny using nuclear ribosomal DNA and cp markers (Wilson et al. 2005; Lucas et al. 2007; Biffin et al. 2010; Thornhill et al. 2015; Berger et al. 2016). Although it has been recently published a phylogenetic work based on complete cp genome sequences from Myrteae tribe (Machado et al. 2017), most of these studies have been performed mainly on Eucalyptus and related genera (Steane 2005; Asif et al. 2013; Bayly et al. 2013; Reginato et al. 2016). Therefore, the availability of complete cp genomes exhibiting new variable and informative regions would help to reconstruct a more accurate phylogeny.

In this study, we present the cp genome of the fleshy fruit, *Eugenia uniflora*, obtained from whole genome sequencing and *de novo* assembly. This represents a solid resource for phylogenetic studies in the Myrtaceae family. We analyzed the genome features of *E. uniflora* and compared them with cp genomes from other Myrtaceae tribes. In addition, we performed a phylogenomic approach using 57 cp genes to reconstruct the phylogeny of Malvidae/Eurosid II group, which includes the Myrtales order.

Materials and methods

Plant material

Young leaves from *Eugenia uniflora* tree were collected from Porto Alegre, RS, Brazil (latitude (S): 30°4′2.71″; longitude (W): 51°7′11.88″). Voucher specimen was deposited at the Herbário do Instituto de Ciências Naturais (ICN 193277).

DNA sequencing and genome assembly

Total DNA was extracted from 1 g of fresh leaves using a CTAB method (Doyle and Doyle 1990). DNA quality was evaluated by electrophoresis on a 1% agarose gel, and quantification was determined using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Total DNA (10 µg) was sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing. One genomic paired-end library of 100-nt-long reads was generated using Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA). To filter reads from the cp genome, the obtained paired-end sequence reads were aligned using Bowtie (Langmead 2010), against *Arabidopsis thaliana* Schur., *Glycine max* Merr., and 40 other Myrtaceae cp genomes (Online Resource 1) with a maximum of two mismatches per read. The filtered reads were assembled with ABYSS software (Simpson et al. 2009). The cp genome scaffolds were orientated by BLAST using the cp genome sequences of *Eucalyptus globulus* Labill and *Eucalyptus grandis* W.Hill as reference genomes (Altschul et al. 1990). Gap regions were filled in after Sanger sequencing using primers F: CATCCGCCAGGAGAGTTTAT, R: AAAGGG CCCTGCTATGAAAA and F: TCGGGTTGTGAGACAC ATTC, R: AACCCGCGTCTTCTCCTT. PCR was carried out in total volume of 20 µl containing 10 ng of DNA, 1× PCR buffer, 1.5 mM $MgCl_2$, 0.25 mM dNTP mix, 0.05 U of Platinum Taq DNA polymerase and 0.5 µM each of forward and reverse primers. The PCR cycle had an initial hot-start step at 94 °C for 5 min, followed by 35 cycles of 94 °C for 45 s, 60 °C for 1 min, 72 °C for 2 min and a final extension step at 72 °C for 5 min. Sanger sequencing reactions were performed using BigDye Terminator v3.1 Cycle sequencing kit and were resolved on ABI 3700 DNA Analyzer.

Genome analysis, codon usage, and repeat structure

Coding sequences (cds), rRNA, and tRNA were annotated using the automatic annotator DOGMA (Dual Organellar GenoME Annotator) (Wyman et al. 2004), verified using

BLAST searches against other plant cp genomes, and finally manually curated. tRNA genes were confirmed by comparison with the appropriate homologs in *Eucalyptus globulus* Labill cp genome and folding-verified with the tRNA scan-SE online program (http://lowelab.ucsc.edu/tRNAscan-SE). The codon usage frequency was analyzed by using MEGA (Tamura et al. 2007). A circular map of the genome was designed using the online OGDRAW program (Lohse et al. 2013). Whole chloroplast gene distribution was performed and visualized between *E. globulus* and *Syzygium cumini* (L.) Skeels. with mVISTA software using *E. uniflora* as the reference genome (Frazer et al. 2004).

The positions and type of simple sequence repeats (SSRs) were detected using MISA (http://pgrc.ipk-gatersleben.de/misa/), with thresholds of eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs. All of the repeats found were manually verified, and redundant results were removed. Tandem repeats were analyzed using Tandem Repeats Finder (TRF) v4.07b (Benson 1999) with the prior mentioned parameter settings. REPuter was used to identify and locate direct and inverted repeats in the cp genome of *E. uniflora* (Kurtz et al. 2001). The minimal repeat size was set to 30 bp, and the identity of repeats was no less than 90% (hamming distance equal to 3).

Phylogenetic analysis

Fifty-seven common cp protein-coding genes (PCGs) (Online Resource 2) were used to infer the phylogenetic relationships among 58 species belonging to the Malvids (Eurosids II) group available in GenBank (Online Resource 3). *Vitis vinifera* L. was set as out-group. Nucleotide sequences were aligned by MUSCLE available in MEGA version 6.0 (Tamura et al. 2007). Phylogenetic trees were generated by the maximum likelihood (ML) method, using the GTR+I+G nucleotide substitution determined by Modeltest ver. 3.7 (Posada and Crandall 1998), using RAxML v8.2.4 (Stamatakis 2014). The stability of each tree node was tested by bootstrap analysis with 1000 replicates. Bayesian analysis on the same dataset was also performed using MrBayes version 3.1.2 (Ronquist and Huelsenbeck 2003). We used the same evolutionary model with 5,000,000 generations sampled every 100 generations. The first 25% of trees were discarded as burn-into produce a consensus phylogram, with posterior probability (PP) values for each node. The phylogenetic trees were rooted and visualized using FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Results

Genome assembly

Reads from Illumina sequencing of the *Eugenia uniflora* nuclear genome were used to assemble the cp genome. The total reads (75,127,218) were filtered and assembled *de novo* into non-redundant contigs and singletons joined into 10 scaffolds. This first draft of the cp genome resulted in mapped reads covering about 99.9% of the genome (coverage 10,938 reads, minimum coverage = 757 reads, maximum coverage = 26 327 reads).

After running BLAST with Eucalyptus genomes, the cp genome sequences resulted in two large scaffolds whose ends were finally closed using PCR and Sanger sequencing. The four junctions between IRs and SSC/LSC were determined by aligning the *E. uniflora* cp genome versus *E. globulus* and *Syzygium cumini* genomes. The final cp genome was then submitted to GenBank (accession number NC_027744).

The overall structure and general features of the *Eugenia uniflora* cp genome

The complete length of the *Eugenia* cp genome is 158,445 bp, and it includes the canonical quadripartite structure consisting of one LSC (87,459 bp), one SSC (18,318 bp) and a pair of IRs (26,334 bp) (Fig. 1). Coding regions (92,848 bp; 58.93%) account for over half of the cp genome, with the peptide-coding regions forming the largest group (81,462 bp; 51.41%), followed by ribosomal RNA genes (9050 bp; 5.71%) and transfer RNA genes (2863 bp; 1.81%). The remaining 41.07% is covered by intergenic regions, introns or pseudogenes (Table 1). The average total AT content is 63% with the IRs having lowest amount (57.2%). A total of 111 different genes, including 30 tRNAs, 4 rRNAs and 77 predicted protein-coding genes, were annotated (Table 2). Among these, seven tRNAs, four rRNAs and six protein-coding genes (*ycf*15, *rps*7, *ndh*B, *ycf*2, *rpl*23, *rpl*2) were present in duplicate in the IR regions. Three pseudogenes, *ycf*1, *ycf*15 *and inf*A, were identified and located in the boundary IRb/SSC, IRb and LSC region, respectively. In the *Eugenia* cp genome, there are 18 gene containing introns, the majority of them (12 genes) are located in the LSC region (four tRNAs and eight protein-coding genes) and the rest are distributed in IRs (two tRNA and three protein-coding genes) and SSC (1 protein-coding gene) region (Table 3). Most of the genes have only one intron, but *clp*P and *ycf*3 have two introns each. The *trn*K$^{(UUU)}$ gene has the largest intron (2530 bp) containing within it the *mat*K gene. The *rps*12 gene sequence is a trans-spliced gene with the 5′end located in

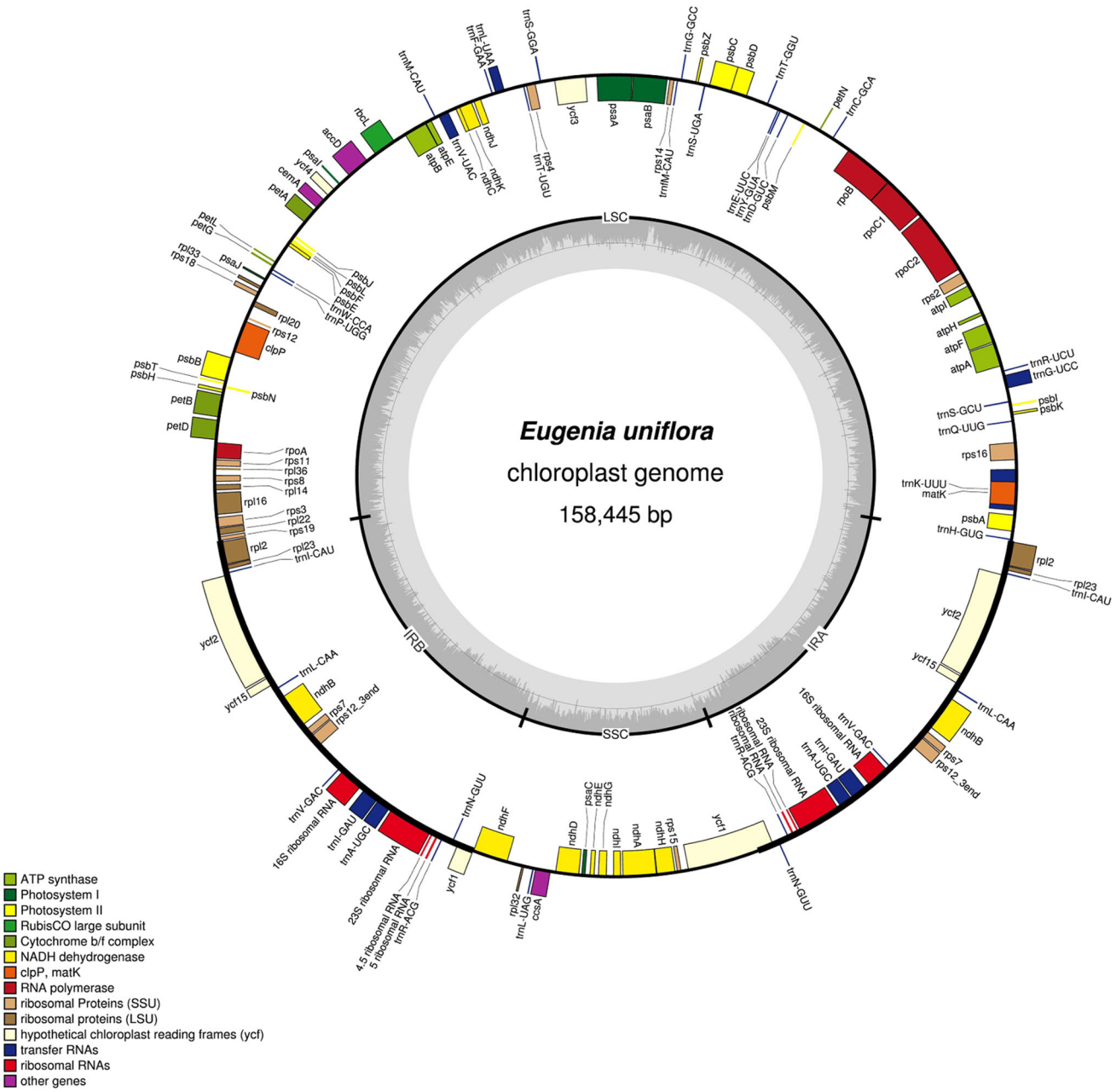


Fig. 1 *Eugenia uniflora* chloroplast genome map. The *thick lines* indicate the extent of the inverted repeats (IRa and IRb), which separate the genome into small and large single-copy regions. Genes on the outside of the map are transcribed clockwise and those on the inside of the map are transcribed counterclockwise. GC content is shown. Gene function or identifiers are displayed by different *colors* as it is indicated by inner legend

the LSC region and the duplicated 3′end in the IR regions. Based on the sequences of protein-coding genes and tRNA genes, the frequency of codon usage was deduced for the cp genome and is summarized in Table 4. The codon usage was biased toward a high representation of A and U at the third codon position, as observed in most land plant cp genes (Ravi et al. 2008).

Comparison of *Eugenia uniflora* to other Myrtaceae cp genomes

The overall sequence alignment of *E. globulus* and *S. cumini* cp genomes was compared using the annotation of *Eugenia uniflora* as a reference. The same order of genes was confirmed because order variations in cp genomes are

Table 1 Summary of the characteristics of *Eugenia uniflora* chloroplast genome

Feature	*E. uniflora*
Total cpDNA size (bp)	158,445
LSC size (bp)	87,459
SSC size (bp)	18,318
IR size (bp)	26,334
Protein-coding regions (%)	58.6%
rRNA and tRNA (%)	7.52%
Introns size (% total)	12.05%
Intergenic sequences and pseudogenes (%)	29.02%
Number of genes	131
Number of different protein-coding genes	77
Number of different tRNA genes	30
Number of different rRNA genes	4
Number of different duplicated genes	17
Pseudogenes	3
GC content	37%

relatively uncommon. The two IRs from the three cp genomes show high similarity in sequence (Fig. 2), on the other hand, the most divergent regions were those localized in the intergenic spacers in the noncoding genes. The coding region sequences show a high level of conservation. Slightly more sequence variation was observed between *E. uniflora* and *E. globulus* cp genomes in the *psa*A, *psa*B and *ycf*2 genes, compared with *S. cumini*.

IR contraction and expansion

In general, *E. uniflora* has the smallest cp genome compared to *E. globulus*, *E. grandis* and *S. cumini* and shows an expansion of the IR over the LSC region (Fig. 3). This also explains the presence of pseudogenes in the border regions, like *ycf*1 in which length variation depends upon if the IR has extended into the SSC region. In the case of *E. uniflora*, a shorter *ycf*1 pseudogene and a larger *ndh*F gene cause a reduction in the intergenic sequence. This last gene is relatively highly variable in the 3′ region (Dong et al.

Table 2 Genes present in *Eugenia uniflora* chloroplast genome

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl*2[b,c], 14, 16[b], 20, 22, 23[c], 32, 33, 36
	Small subunit of ribosomal proteins	*rps*2, 3, 4, 7[c], 8, 11, 12[b–d], 14, 15, 16[b], 18, 19
	rRNA genes	*rrn*4.5, 5, 16, 23
	tRNA genes	*trn*A[(UGC)b,c], C[(GCA)], D[(GUC)], E[(UUC)], F[(GAA)], G[(UCC)b,c], G[(GCC)], H[(GUG)], I[(CAU)c], I[(GAU)b,c], K[(UUU)b], L[(UAG)], L[(CAA)c], L[(UAA)b], M[(CAU)], fM[(CAU)], N[(GUU)c], Q[(UUG)], R[(ACG)c], R[(UCU)], S[(GGA)], S[(GCU)], S[(UGA)], T[(GGU)], T[(UGU)], V[(UAC)b], V[(GAC)c], W[(CCA)], Y[(GUA)], P[(UGG)]
Photosynthesis	Photosystem I	*psa*A, B, C, I, J, *ycf*3[a], *ycf*4
	Photosystem II	*psb*A, B, C, D, E, F, H, I, J, K, L, M, N, T, Z
	NADH oxidoreductase	*ndh*A[b], B[b,c], C, D, E, F, G, H, I, J, K
	Cytochrome b6/f complex	*pet*A, B[b], D[b], G, L, N
	ATP synthase	*atp*A, B, E, F[b], H, I, L
	Rubisco	*rbc*L
Other gene	Maturase	*mat*K
	Protease	*clp*P[a]
	Envelop membrane protein	*cem*A
	Subunit Acetyl-CoA carboxylase	*acc*D
	c-type cytochrome synthesis gene	*ccs*A
Unknown gene	Conserved open reading frames	*ycf*1, *ycf*2[c], *ycf*15[c]

[a] Genes containing two introns

[b] Genes containing a single intron

[c] Genes with two copies

[d] Genes split into two independent transcription units

Table 3 Genes with introns in the *Eugenia uniflora* chloroplast genome and the length of the exons and introns

Gene	Location	exon I (bp)	intron I (bp)	exon II (bp)	intron II (bp)	exon III (bp)
*trn*K$^{(UUU)}$	LSC	37	2568	35		
*rps*16	LSC	39	867	204		
*trn*G$^{(UCC)}$	LSC	23	755	49		
*atp*F	LSC	147	742	408		
*rpo*C1	LSC	453	729	1614		
*ycf*3	LSC	126	758	228	727	148
*trn*L$^{(UAA)}$	LSC	37	502	46		
*trn*V$^{(UAC)}$	LSC	39	600	37		
*clp*P	LSC	69	866	291	619	223
*pet*B	LSC	6	771	639		
*pet*D	LSC	9	752	471		
*rpl*16	LSC	9	1000	396		
*rpl*2	IR	390	664	432		
*ndh*B	IR	777	681	753		
*rps*12	IR	210	567	27		
*trn*I$^{(GAU)}$	IR	37	957	35		
*trn*A$^{(UGC)}$	IR	38	803	35		
*ndh*A	SSC	549	1067	537		

2012). The intergenic spacer located between the IRA/LSC border and the *trn*H gene of the LSC region established differences between the cp genomes. This region is 43 bp in *E. uniflora,* similar to that of *S. cumini* (55 bp), but different from other dicots where it ranges in size of 2–12 bp (Shinozaki et al. 1986; Ibrahim et al. 2006).

Repeat structure and SSR analysis

For repeat structure analysis, eleven forward, one inverted, and twelve tandem repeats were detected in the *E. uniflora* cp genome (Table 5). Most of these repeats (67%) exhibited lengths between 20 and 50 bp and were located in intergenic spacers regions and introns. The coding regions of *psa*A, *psa*B, *ycf*1 and *ycf*2 genes showed some repeated sequences. Although the number of repeats was variable respect to *Syzygium* and *Eucalyptus*, they were identified in the same genes. Most of the repeated regions identified in this work have already been compared in *S. cumini, Eugenia grandis, E. globulus, Nicotiana tabacum* L., *Gossypium barbadense* L. and show a high degree of conservation (Asif et al. 2013). It appears that dispersed repeats are very common in angiosperm cp genomes, but future comparative studies are needed to determine the functional and evolutionary role of these repeats.

SSRs are repeated DNA sequences consisting of direct tandem repeats of short (1–10 bp) nucleotide motifs. In this study, a total of 215 SSR loci were identified, most of them (76.25%) were A and T mononucleotide repeats (Table 6) similar to other Myrtaceae cp genomes (Asif et al. 2013). Most SSRs are located in intergenic regions, but some were found in *ndh*F, *pet*A, *ycf*2, *rpo*C2, *psa*J, *psb*B, *ycf*1, *ccs*A, *ycf*4 and *rps*19 coding genes (Table 6).

Phylogenetic analysis

In this study, the concatenated nucleotide sequences of 57 PCGs of 58 cp genomes of Malvidae group were used to reconstruct the phylogenetic relationships by the ML and Bayesian method. These 57 genes were present in all the cp genomes so the problem of missing data from the sequence alignment was minimized. The sequence alignment used comprised 36,206 characters. The final alignment was submitted and assigned as 21,047 in the TreeBASE database (https://treebase.org/). ML analysis resulted in a single tree with ln $L = -249{,}032.011$, and bootstrap values were high with values >80% for 4 of 55 nodes, and 48 nodes with 100% bootstrap (Online Resource 4). Although the Bayesian and ML analyses showed similar topologies, the posterior probabilities in the Bayesian analysis were better than the bootstrap values in the ML (Fig. 4). Therefore, only the Bayesian tree was chosen for discussing the phylogenetic results.

There are congruence areas strongly supported by the phylogeny (PP = 1.0) that include the monophyly of Brassicales and their sister relationship to Malvales and Sapindales and monophyly of Geraniales and Myrtales. Our phylogenies placed Myrtales in a sister relationship to Geraniales with solid support and resolution (PP = 0.95),

Table 4 Codon–anticodon recognition pattern and codon usage for the *Eugenia uniflora* chloroplast genome

Codon	Aminoacid	Count	RSCU	*trn*A	Codon	Aminoacid	Count	RSCU	*trn*A
UUU	F	2308	1.19	*trn*F$^{(GAA)}$	UAU	Y	1456	1.34	*trn*Y$^{(GUA)}$
UUC	F	1587	0.81		UAC	Y	715	0.66	
UUA	L	1080	1.19	*trn*L$^{(UAA)}$	UAA	*	1225	1.21	
UUG	L	1160	1.28	*trn*L$^{(CAA)}$	UAG	*	855	0.84	
CUU	L	1110	1.22	*trn*L$^{(UAG)}$	CAU	H	967	1.4	*trn*H$^{(GUG)}$
CUC	L	717	0.79		CAC	H	416	0.6	
CUA	L	848	0.94		CAA	Q	1102	1.39	*trn*Q$^{(UUG)}$
CUG	L	526	0.58		CAG	Q	478	0.61	
AUU	I	1888	1.21	*trn*I$^{(GAU)}$	AAU	N	1819	1.39	*trn*N$^{(GUU)}$
AUC	I	1230	0.79		AAC	N	795	0.61	
AUA	I	1565	1	*trn*I$^{(CAU)}$	AAA	K	2172	1.32	*trn*K$^{(UUU)}$
AUG	M	958	1	*trn*(f)M$^{(CAU)}$	AAG	K	1117	0.68	
GUU	V	839	1.36	*trn*V$^{(GAC)}$	GAU	D	1025	1.41	*trn*D$^{(GUC)}$
GUC	V	437	0.71		GAC	D	429	0.59	
GUG	V	446	1.22		GAA	E	1379	1.38	*trn*E$^{(UUC)}$
GUA	V	754	0.72	*trn*V$^{(UAC)}$	GAG	E	622	0.62	
UCU	S	1117	1.48	*trn*S$^{(GGA)}$	UGU	C	667	1.2	*trn*C$^{(GCA)}$
UCC	S	855	1.13		UGC	C	449	0.8	
UCG	S	602	0.8		UGA	*	956	0.94	
UCA	S	861	1.14	*trn*S$^{(UGA)}$	UGG	W	704	1	*trn*W$^{(CCA)}$
CCU	P	662	1.07	*trn*P$^{(UGG)}$	CGU	R	321	0.6	*trn*R$^{(ACG)}$
CCC	P	564	0.92		CGC	R	252	0.47	*trn*R$^{(UCU)}$
CCA	P	799	1.3		CGA	R	577	1.08	
CCG	P	440	0.71		CGG	R	363	0.68	
ACU	T	647	1.17	*trn*T$^{(GGU)}$	AGA	R	1079	2.01	
ACC	T	549	0.99		AGG	R	626	1.17	
ACG	T	362	0.65		AGU	S	627	0.83	*trn*S$^{(GCU)}$
ACA	T	656	1.19	*trn*T$^{(UGU)}$	AGC	S	471	0.62	
GCU	A	469	1.3	*trn*A$^{(UGC)}$	GGU	G	510	0.95	*trn*G$^{(GCC)}$
GCC	A	329	0.91		GGC	G	330	0.62	
GCA	A	420	1.16		GGG	G	537	1	
GCG	A	225	0.62		GGA	G	764	1.43	*trn*G$^{(UCC)}$

despite the fact that this order still has a controversial position in respect to other members of the Rosids (Fig. 4).

In analyzing the Myrtales clade, we showed a closer relationship between species from Melastomataceae and Myrtaceae family than to Onagraceae family. Our phylogenetic tree clearly supports the monophyly of the three Myrtoideae tribes: Myrteae, Eucalypteae and Syzygieae (PP = 1.0). Additionally, we corroborated the paraphyly of Corymbia in the Eucalypteae tribe and observed that the latter has a closer relationship to Syzygieae than Myrteae (Bayly et al. 2013). *Eugenia uniflora* is placed along with *Acca sellowiana* (O.Berg) Burret as the diverging lineage, and they have a closer relationship with *S. cumini* (Syzygieae tribe) than to the Eucalypteae tribe.

Discussion

The cp genome of *Eugenia uniflora* was assembled *de novo* from the Illumina NGS reads derived from the whole genome. This approach, without prior purification of the cpDNA, provides a new way to obtain the cp genome and has been successful in several studies (Leseberg and Duvall 2009; Tangphatsornruang et al. 2010; Straub et al. 2011). Our work serves as another example of this approach for obtaining high coverage (99%) of the cp genome.

The *E. uniflora* cp genome has the typical quadripartite structure (Fig. 1) and gene content with a size in range with other Myrtaceae family members (Asif et al. 2013; Bayly et al. 2013; Machado et al. 2017). Major differences among angiosperm cp genomes are due to gene loss, inversions,

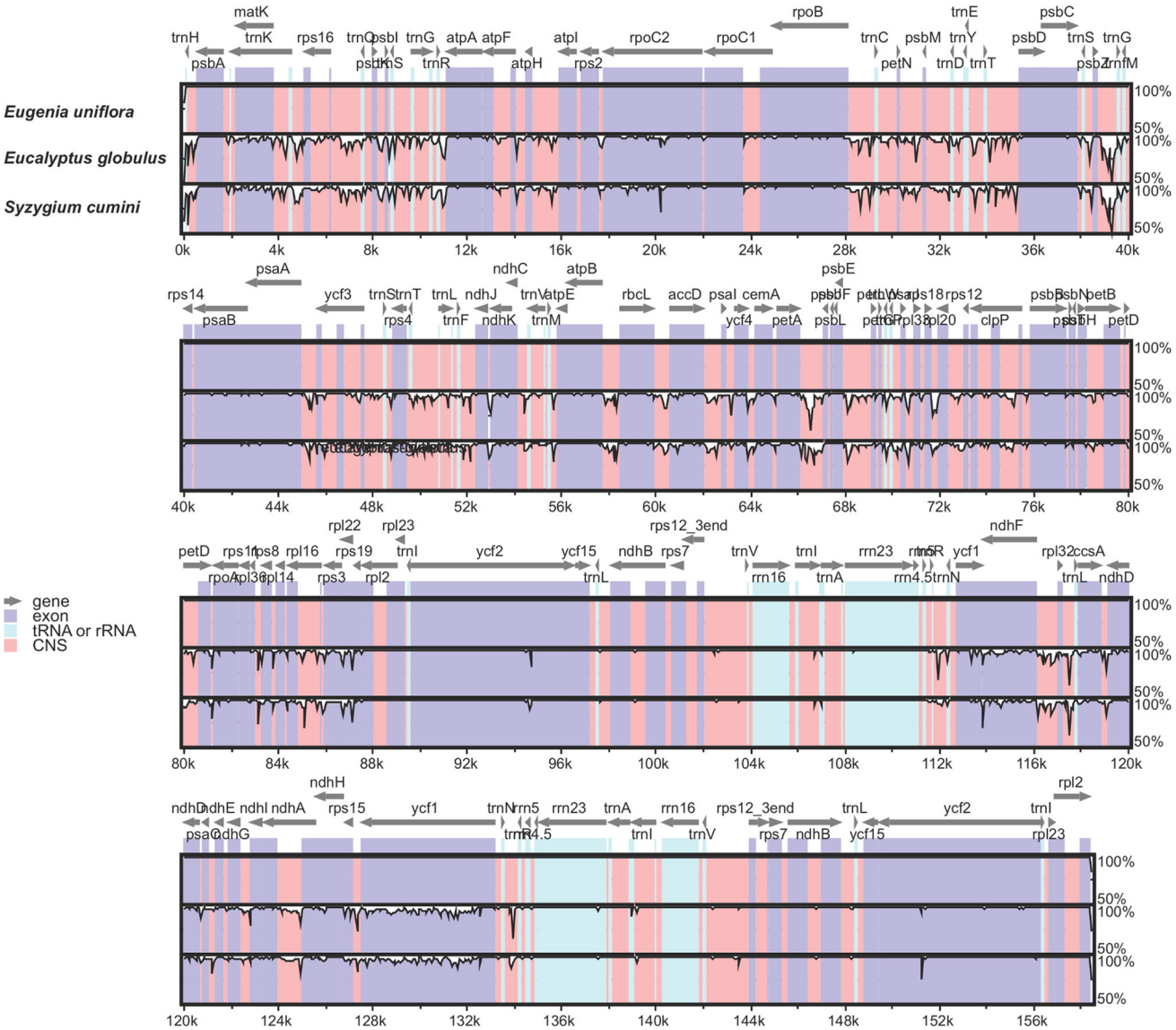


Fig. 2 Sequence identity *plot* comparing the chloroplast genome of *Eugenia uniflora* to other Myrtaceae. Pairwise comparisons between *E. uniflora* and *E. globulus* (*top*) and *Syzygium cumini* (*bottom*) chloroplast genomes using mVISTA. The *y*-axis represents % identity ranging from 50 to 100%. Coding, rRNA, tRNA and conserved noncoding sequences (CNS) are shown as indicated by inner legend

and expansion/contraction of inverted repeat regions. The IR contraction and expansion events, the presence of many stop codons in the coding sequence, or a probable partial duplication are all reasons that could explain the presence of pseudogenes in the cp genome. In our work, this is represented in the *ycf*1, *inf*A and *ycf*15 pseudogenes (Fig. 1; Table 1). Some alternative codons were also identified, ACG was used as an alternative initiation codon in the *psb*L and *ndh*D genes and GUG was only found as a start codon in the *ycf*15 pseudogene and *rps*19 gene. ACG has been shown to convert to an AUG initiation site as reported in *N. tabacum* (Sasaki 2003), and GUG codons have been reported to be more efficient than ACG in translation initiation (Rohde et al. 1994). Most cp genomes are quite AT rich with (above 60%) unevenly distributed AT contents, as well as conserved regions of lower AT contents. The features of the *E. uniflora* cp genome are the same, and of all the cp regions, the IRs have the lowest AT content (57.2%) because of the presence of ribosomal genes (Ravi et al. 2008). These values are congruent with those reported in other Myrtaceae cp genomes (Asif et al. 2013; Bayly et al. 2013; Machado et al. 2017).

Chloroplast SSRs (cpSSRs) are generally short mononucleotide tandem repeats that, when located in the noncoding regions of the cp genome, commonly show intraspecific variation in repeat number. CpSSRs can exhibit high variation within the same species and thus are considered valuable markers for population genetics

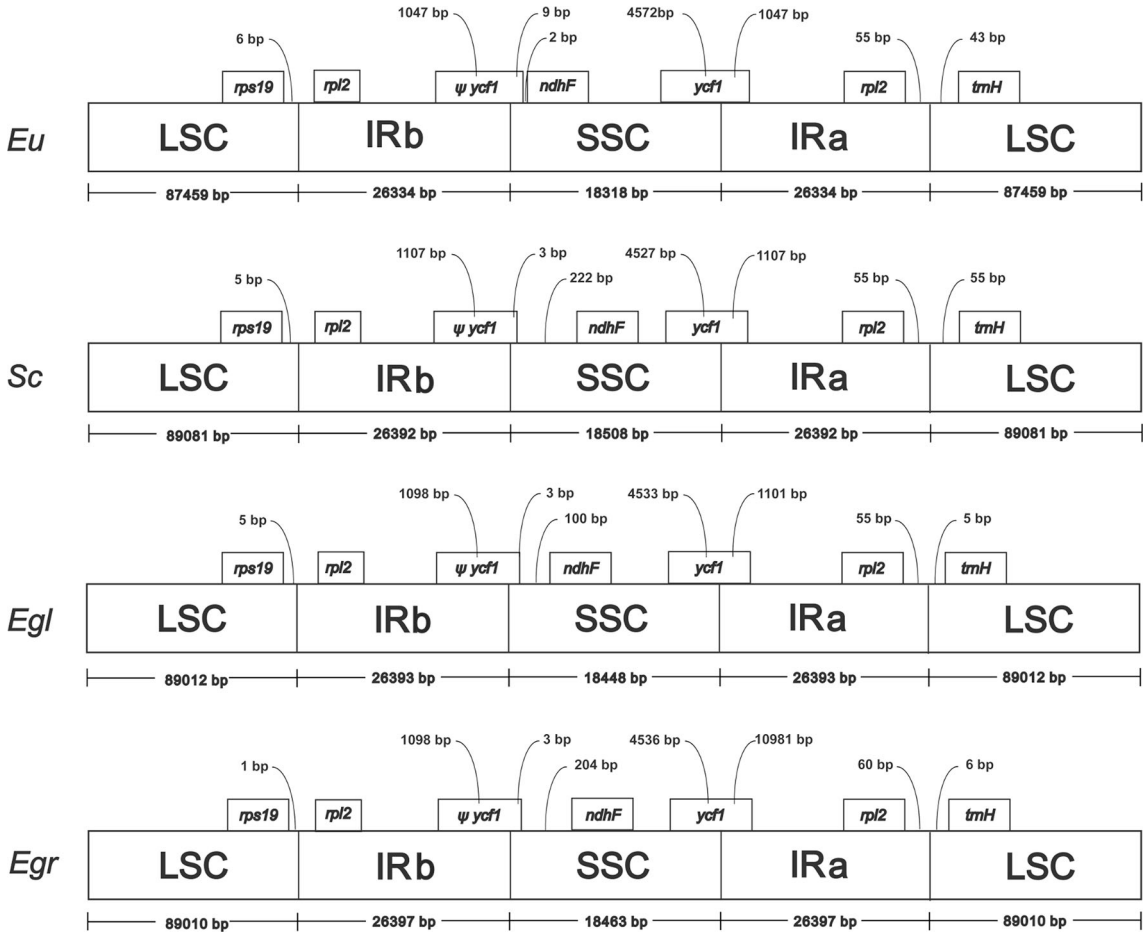


Fig. 3 Comparison of border positions of LSC, SSC and IR among *Eugenia uniflora* and related Myrtaceae family species. *Boxes* above the *main line* indicate the predicted genes, while pseudogenes at the borders are shown by Ψ. Their length is displayed in the corresponding regions. The figure is not scaled and just shows relative changes at or near the IR-SC borders. Sc *Syzygium cumini*, Egl *E. globulus*, Egr *E. grandis*

(Provan et al. 2001). In this work, we identified some SSRs that can be utilized to increase our understanding of the genetic structure of *E. uniflora* populations (Margis et al. 2002; Salgueiro et al. 2004; Ferreira-Ramos et al. 2008). Understanding the effects of spatial isolation on the levels of genetic diversity and gene flow is crucial to providing recommendations for in situ and ex situ conservation of the species. In addition, these SSR markers will also be useful in future studies of other Myrtaceae species from the Neotropics.

Although previous phylogenetic studies improved our understanding of intergeneric relationships within the Myrtales order, the relationship between fleshy-fruited and dry-capsular clades remains unresolved. In this work, some representative cp genomes from Melastomataceae, Onagraceae and Myrtaceae family were selected to build a Malvidae metatree. We used species from the Malvidae group because the order Myrtales belongs to this group and there are several cp genomes available. To do this, 57 protein-coding genes for 13 taxa were analyzed using both the ML and Bayesian methods. Both trees are congruent to that presented in a recent study using 78 cp coding genes from 30 angiosperm taxa (Ruhfel et al. 2014) and to that using 72 complete cp genomes from Rosids (Su et al. 2014). Although our results clearly favor a closer relationship of Myrtales to the Geraniales clade, expanded sampling of complete cp genome sequences of Rosids is needed to resolve this issue, especially since limited taxon sampling can lead to erroneous tree topologies (Leebens-Mack et al. 2005).

Eugenia uniflora formed one monophyletic clade along with *A. sellowiana*, another Myrtaceae from Neotropical region, as previously reported by Machado et al. 2017 using complete cp genomes. These two species were more closely related to *Syzygium cumini* than the Eucalypteae tribe. The Syzygieae tribe has had a long association with the predominantly New World Myrtaceous, mostly because they showed a high similarity between their cp complete

Table 5 Repeated sequences in the *Eugenia uniflora* chloroplast genome

Repeat size (bp)	Start position of first repeat	Type[a]	Start position the repeat found in other region	Copy number	Location[b]	Region
15	55,610	T	55,625	(×2)	IGS (*trn*M$^{(CAU)}$-*atp*E)	LSC
15	130,810	T	130,825, 130,840	(×3)	*ycf*1	SSC, IRB
16	10,031	T	10,047	(×2)	intron (*trn*S$^{(UCC)}$)	LSC
16	87,134	T	87,150	(×2)	IGS (*rpl*22-*rps*19)	LSC
17	102,403	T	102,420	(×2)	IGS (*rps*12-*trn*V$^{(GAC)}$)	IRA
18	66,388	T	66,406	(×2)	IGS (*pet*A-*psb*J)	LSC
18	94,667	T	94,685, 94,703	(×3)	*ycf*2	IRA
20	6373	T	6393	(×2)	IGS (*rps*16-*trn*Q$^{(UUG)}$)	LSC
20	39,036	T	39,056	(×2)	IGS (*psb*Z-*trn*G$^{(GCC)}$)	LSC
20	70,664	T	70,684	(×2)	IGS (*psa*J-*rpl*33)	LSC
21	92,239	T	92,260, 92,281	(×3)	*ycf*2	IRA
31	92,238	F	92,259	(×2)	*ycf*2	IRA
31	45,820	F	102,036	(×2)	intron I (*ycf*3); IGS (*rps*12-*trn*V$^{(GAC)}$)	LSC, IRA
31	153,617	F	153,638	(×2)	*ycf*2	IRB
32	8762	F	38,011	(×2)	IGS (*psb*I-*trn*S$^{(GCU)}$, *trn*S$^{(GCU)}$; IGS (*psb*C-*trn*S$^{(UGA)}$, *trn*S$^{(UGA)}$)	LSC
35	70,665	I	70,665	(×2)	IGS (*psa*J-*rpl*33)	LSC
39	46,761	F	102,014	(×2)	intron II (*ycf*3); IGS (*rps*12-*trn*V$^{(GAC)}$)	LSC, IRA
40	102,014	F	123,971	(×2)	*rps*12; intron (*ndh*A)	IRA, SSC
41	41,804	F	44,028	(×2)	*psa*B; *psa*A	LSC
42	46,758	F	123,968	(×2)	intron II (*ycf*3); intron *ndh*A	LSC, SSC
45	94,666	F	94,684	(×2)	*ycf*2	IRA
45	151,175	F	151,193	(×2)	*ycf*2	IRB
50	38,352	T	38,402	(×2)	IGS (*trn*S$^{(UGA)}$-*psb*Z)	LSC
62	38,351	F	38,401	(×2)	IGS (*trn*S$^{(UGA)}$-*psb*Z)	LSC

[a] *F* Forward; *I* Inverted; *T* Tandem

[b] *IGS* intergenic spacer region

genome sequences. Additionally, they have characteristics in common such as fleshy large-seeded fruits, biotic dispersal, and they are both woody rainforest trees. These results are in agreement with Biffin et al. (2010), who concluded that Syzygieae and Myrteae show highly significant positive variation in diversification rates associated with both of these lineages relative to the overall evolutionary radiation of Myrtaceae. Our phylogenetic tree also confirmed the closer relationship between Melastomataceae and Myrtaceae than to the Onagraceae family as reported by previous analyses based on complete cpDNA (Berger et al. 2016; Reginato et al. 2016). Our phylogenetic analyses based on complete cp genomes further expand the taxon sampling of entire genomes as we included one more Neotropical Myrtaceae genome in a metatree analysis.

Conclusions

The *Eugenia uniflora* cp genome organization and gene content are typical of most angiosperms and are similar to that of Myrtaceae species. It features a relevant number of simple sequence repeats, which could be further explored for population studies within the *Eugenia* genus. Moreover, these data increase the genetic and genomic resources available in Myrtaceae by adding a new strategy of

Table 6 List of simple sequence repeats in *Eugenia uniflora*. The SSR-containing coding regions are indicated in parentheses

Repeat unit	Length (bp)	Number of SSRs	Start position
A	8	31	1992; 4532; 6547; 6772; 8905; 9344; 14,290; 19,917; 23,570; 23,757; 39,529; 45,575; 45,600; 55,722; 62,539; 65,559 (*pet*A); 68,727; 68,795; 72,516; 75,695; 81,223; 113,806 (*ndh*F); 114,538 (*ndh*F); 118,402; 120,673; 120,744; 131,958; 139,500; 143,496; 146,803; 158,425
	9	22	17; 7940; 9048; 12,656; 13,443; 14,472; 31,941; 32,744; 32,817; 38,913; 47,458; 48,054; 57,737; 71,263; 71,766; 92,859 (*ycf*2); 116,745; 117,617; 118,856; 122,698; 126,805; 134,276
	10	10	303; 4705; 4780; 8176; 47,420; 48,211; 48,271; 62,286; 74,562; 131,613
	11	7	5678; 8732; 50,518; 75,032; 117,327; 124,318; 130,258
	12	2	60,317; 84,317
	13	1	8707; 74,079
	15	1	14,765
	19	1	32,343
T	8	33	4296; 5792; 8338; 18,391 (*rpo*C2); 29,269; 31,642; 37,906; 45,860; 69,550; 70,515 (*psa*J); 70,892; 74,164; 76,564 (*psb*B); 78,464; 84,363; 85,593; 85,682; 87,473; 99,095; 102,402; 106,398; 117,953 (*ccs*A); 118,488 (*ccs*A); 119,042; 119,067; 119,739; 127,133; 127,935 (*ycf*1); 128,476 (*ycf*1); 130,326 (*ycf*1); 131,338 (*ycf*1); 131,455 (*ycf*1); 131,573 (*ycf*1)
	9	22	141; 2481; 9565; 14,030; 19,668; 31,404; 34,703; 47,358; 49,898; 54,438; 62,923; 68,293; 74,663; 87,435 (*rps*19); 111,621; 117,239; 122,789; 124,393; 128,707; 130,023 *(ycf*1); 130,846 (*ycf*1); 153,038 (*ycf*2)
	10	17	7863; 9093; 10,950; 15,525; 22,370 (*rpo*C1 - exon II); 27,435 (*rpo*B); 45,944 (*ycf*3 - intron I); 47,321 (*ycf*3 - intron II); 54,299; 54,696; 57,631; 70,066; 72,927; 74,602; 75,767; 85,868; 86,670 (*rpl*22)
	11	8	13,215; 17,551; 19,774 (*rpo*C2); 63,472 (*ycf*4); 66,969; 73,028; 73,341; 124,796
	12	2	66,289; 73,896
	13	1	70,607
	14	2	15,692; 53,692 (*ndh*K)
	15	1	83,775
	20	1	51,433
C	8	2	39,678; 65,485 (*pet*A)
AG	8	2	98,454 (*ndh*B - exon I); 136,156 (*rrn*23)
AT	8	15	1884; 10,527; 45,420; 58,748 (*rbc*L); 60,365; 60,817 (*acc*D); 62,673; 65,199 (*pet*A); 66,816; 70,297; 87,039 (*rpl*22); 124,099; 127,519; 148,203; 157,836
	10	1	33,835
CA	8	1	3100
CT	8	3	31,961; 109,742 (*rrn*23); 147,444 (*ndh*B - exon II)
GA	8	4	38,017 ($trnS^{(UGA)}$; 58,932 (*rbc*L); 90,677 (*ycf*2); 92,880 (*ycf*2)
TA	8	5	7506; 88,061; 96,251 (*ycf*2); 97,694; 149,647 (*ycf*2)
TC	8	3	131,255 (*ycf*1); 153,018 (*ycf*2); 155,221 (*ycf*2)
	10	1	64,285 (*cem*A)
AGA	12	1	139,167
CAG	12	1	1177 (*psb*A)
TTA	12	1	68,856
TTC	12	1	106,726
AATA	12	1	119,348 (*ndh*D)
AGAT	12	1	4894
ATAG	12	1	115,884 (*ndh*F)
ATTA	12	1	33,664
ATTT	12	1	11,090
CTTG	12	1	29,446
TAAG	12	1	46,202
TAAT	12	1	129,206 (*ycf*1)
TCTT	12	1	63,902
TTAT	12	1	78,171
TTTC	12	2	78,202; 85,555

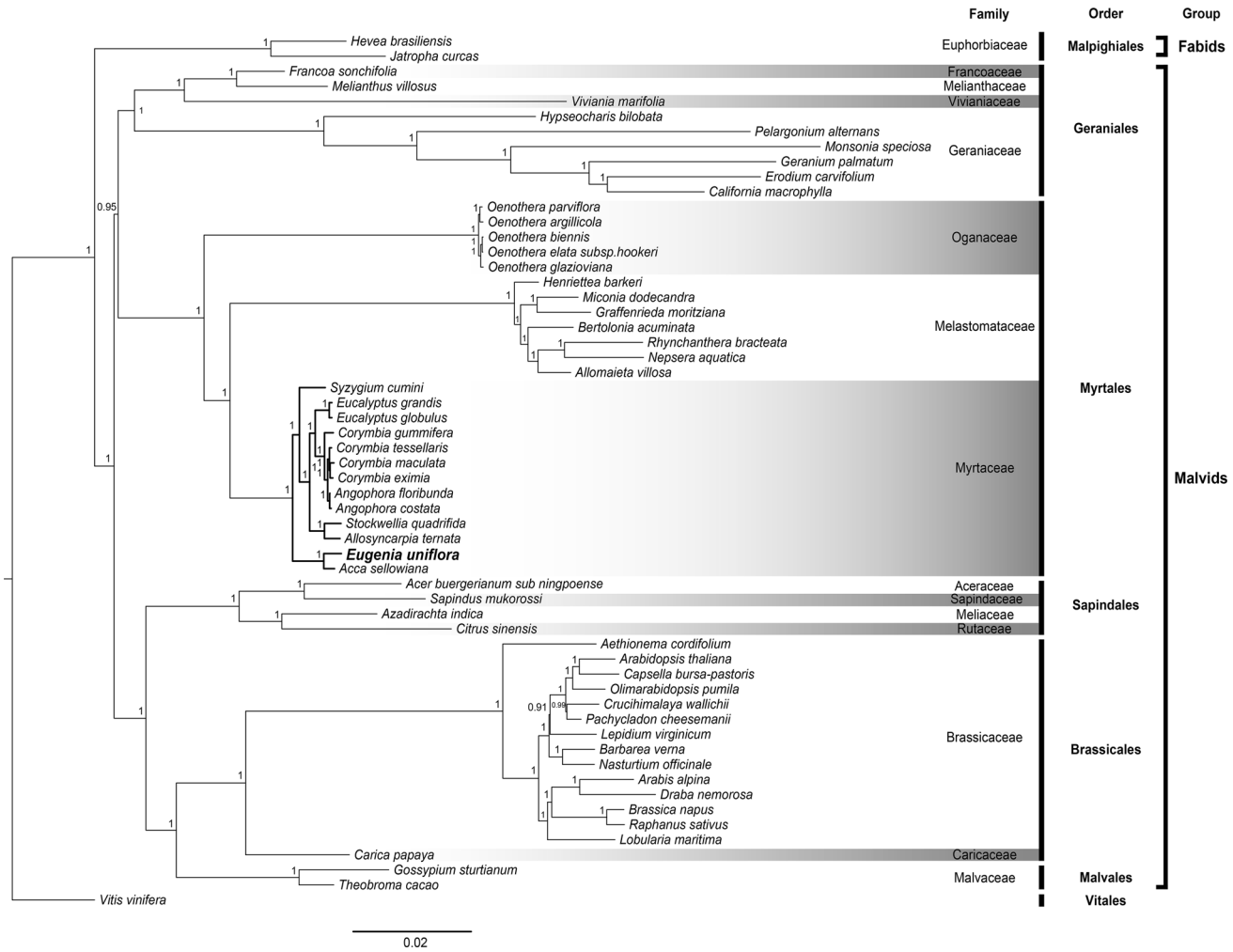


Fig. 4 Phylogenetic tree among 58 eurosids based on 57 protein-coding genes reconstructed by the Bayesian method. The posteriori probabilities are labeled at nodes. Family, order and higher-level group names are also indicated

organelle genome assembly. The cp genome reported here will enrich and help to resolve the phylogeny of the Rosids subclass. In addition, studies of the *Eugenia uniflora* genome will also allow for discovery and interpretation of functional elements encoded within those sequences, providing a basis for understanding key evolutionary changes that correlate with the high diversification rate of Myrteae tribe.

Acknowledgements We would like to thank Prof. Andreia Turchetto Zolet for providing very helpful suggestions to our manuscript and Steven Clipman for correcting the English.

Funding This study was carried out with the support of FAPERGS and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Information on Electronic Supplementary material

Online Resource 1 List of accession numbers of the chloroplast genome sequences used to filter reads from cp genomes.
Online Resource 2 List of the 57 genes used in the ML and Bayesian phylogenetic analysis.
Online Resource 3 List of plastome sequences of Rosids included in the ML and Bayesian phylogenetic analyses.
Online Resource 4 Phylogenetic tree among 58 eurosids based on 57 protein-coding genes reconstructed by the maximum likelihood method.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Molec Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Asif H, Khan A, Iqbal A, Khan IA, Heinze B, Azim MK (2013) The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. Tree Genet Genomes 9:867–877. doi:10.1007/s11295-013-0604-1

Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, Bossinger G, Merchant A, Udovicic F, Woodrow IE, Tibbits J (2013) Chloroplast genome analysis of Australian eucalypts—*Eucalyptus, Corymbia, Angophora, Allosyncarpia* and *Stockwellia* (Myrtaceae). Molec Phylogen Evol 69:704–716. doi:10.1016/j.ympev.2013.07.006

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucl Acids Res 27:573–580. doi:10.1093/nar/27.2.573

Berger BA, Kriebel R, Spalink D, Sytsma KJ (2016) Divergence times, historical biogeography, and shifts in speciation rates of Myrtales. Molec Phylogen Evol 95:116–136. doi:10.1016/j.ympev.2015.10.001

Biffin E, Lucas EJ, Craven LA, Da Costa IR, Harrington MG, Crisp MD (2010) Evolution of exceptional species richness among lineages of fleshy-fruited Myrtaceae. Ann Bot (Oxford) 106:79–93. doi:10.1093/aob/mcq088

De Almeida DJ, Faria MV, Da Silva PR (2012) Biologia experimental em Pitangueira: uma revisão de cinco décadas de publicações científicas/Experimental biology in pitangueira: a review of five decades of scientific publications. Rev Ambiência 8:159–175. doi:10.5777/ambiencia.2012.01.02rb

Dong W, Liu J, Yu J, Wang L, Zhou S (2012) Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLOS One 7:e35071. doi:10.1371/journal.pone.0035071

Doyle J, Doyle J (1990) Isolation of plant DNA from fresh tissue. Focus (Madison) 12:13–15

Ferreira-Ramos R, Laborda PR, De Oliveira Santos M, Mayor MS, Mestriner MA, De Souza AP, Alzate-Marin AL (2008) Genetic analysis of forest species *Eugenia uniflora* L. through of newly developed SSR markers. Conservation Genet 9:1281–1285. doi:10.1007/s10592-007-9458-0

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucl Acids Res 32(Web Server issue): W273–W279. doi:10.1093/nar/gkh458

Govaerts R, Sobral M, Ashton P, Barrie F, Holst BK, Landrum LL, Matsumoto K, Mazine FF, Lughadha EN, Proenca C, Soares-Silva LH, Wilson PG, Lucas E (2015) World checklist of Myrtaceae. Royal Botanic Gardens, Kew

Ibrahim RIH, Azuma J-I, Sakamoto M (2006) Complete nucleotide sequence of the cotton (*Gossypium barbadense* L.) chloroplast genome with a comparative analysis of sequences among 9 dicot plants. Genes Genet Syst 81:311–321

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee S-B, Peery R, McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374. doi:10.1073/pnas.0709121104

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl Acids Res 29:4633–4642. doi:10.1093/nar/29.22.4633

Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinf 32:11.7.1–11.7.14. doi:10.1002/0471250953.bi1107s32

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Molec Biol Evol 22:1948–1963. doi:10.1093/molbev/msi191

Leseberg CH, Duvall MR (2009) The complete chloroplast genome of coix lacryma-jobi and a comparative molecular evolutionary analysis of plastomes in cereals. J Molec Evol 69:311–318. doi:10.1007/s00239-009-9275-9

Lim TK (2012) *Eugenia uniflora*. In: Lim TK, Edible Medicinal and Non Medicinal Plants, vol. 3, Fruits. Springer, Netherlands, pp 620–630. doi:10.1007/978-94-007-2534-8_85

Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucl Acids Res 41(Web Server issue):W575–W581. doi:10.1093/nar/gkt289

Lucas EJ, Harris SA, Mazine FF, Belsham SR, Nic Lughadha EM, Telford A, Gasson PE, Chase MW (2007) Suprageneric phylogenetics of Myrteae, the generically richest tribe in Myrtaceae (Myrtales). Taxon 56:1105–1128. doi:10.2307/25065906

Machado LO, Vieira LD, Stefenon VM, Pedrosa OF, De Souza EM, Guerra MP, Nodari RO (2017) Phylogenomic relationship of feijoa (*Acca sellowiana* (O.Berg) Burret) with other Myrtaceae based on complete chloroplast genome sequences. Genetica 145:1–12. doi:10.1007/s10709-017-9954-1

Margis R, Felix D, Caldas JF, Salgueiro F, De Araujo DSD, Breyne P, Van Montagu M, De Oliveira D, Margis-Pinheiro M (2002) Genetic differentiation among three neighboring Brazil-cherry (*Eugenia uniflora* L.) populations within the Brazilian Atlantic rain forest. Biodivers & Conservation 11:149–163. doi:10.1023/A:1014028026273

Mazine FF, Souza VC, Sobral M, Forest F, Lucas E (2014) A preliminary phylogenetic analysis of *Eugenia* (Myrtaceae: Myrteae), with a focus on Neotropical species. Kew Bull 69:1–14. doi:10.1007/s12225-014-9497-x

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci USA 107:4623–4628. doi:10.1073/pnas.0907801107

Pennington RT, Lavin M, Oliveira-Filho A (2009) Woody Plant Diversity, Evolution, and Ecology in the Tropics: perspectives from Seasonally Dry Tropical Forests. Annual Rev Ecol Evol Syst 40:437–457. doi:10.1146/annurev.ecolsys.110308.120327

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818. doi:10.1093/bioinformatics/14.9.817

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147. doi:10.1016/S0169-5347(00)02097-8

Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. Pl Syst Evol 271:101–122. doi:10.1007/s00606-007-0608-0

Reginato M, Neubig KM, Majure LC, Michelangeli FA (2016) The first complete plastid genomes of Melastomataceae are highly structurally conserved. PeerJ 4:e2715. doi:10.7717/peerj.2715

Rohde W, Gramstat A, Schmitz J, Tacke E, Prufer D (1994) Plant viruses as model systems for the study of non-canonical translation mechanisms in higher plants. J Gen Virol 75:2141–2149. doi:10.1099/0022-1317-75-9-2141

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. doi:10.1093/bioinformatics/btg180

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG (2014) From algae to angiosperms - inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol 14:23. doi:10.1186/1471-2148-14-23

Salgueiro F, Felix D, Caldas JF, Margis-Pinheiro M, Margis R (2004) Even population differentiation for maternal and biparental gene markers in *Eugenia uniflora*, a widely distributed species from the Brazilian coastal Atlantic rain forest. Diversity & Distrib 10:201–210. doi:10.1111/j.1366-9516.2004.00078.x

Sasaki T (2003) Identification of RNA editing sites in chloroplast transcripts from the maternal and paternal progenitors of tobacco (*Nicotiana tabacum*): comparative analysis shows the involvement of distinct trans-factors for ndhB editing. Molec Biol Evol 20:1028–1035. doi:10.1093/molbev/msg098

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123. doi:10.1101/gr.089532.108

Spada PDS, De Souza GGN, Bortolini GV, Henriques JAP, Salvador M (2008) Antioxidant, mutagenic, and antimutagenic activity of frozen fruits. J Med Food 11:144–151. doi:10.1089/jmf.2007.598

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. doi:10.1093/bioinformatics/btu033

Steane DA (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). DNA Res 12:215–220. doi:10.1093/dnares/dsi006

Straub SCK, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A (2011) Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. BMC Genomics 12:211. doi:10.1186/1471-2164-12-211

Su H-J, Hogenhout SA, Al-Sadi AM, Kuo C-H (2014) Complete Chloroplast Genome Sequence of Omani Lime (*Citrus aurantiifolia*) and Comparative Analysis within the Rosids. PLOS One 9:e113049. doi:10.1371/journal.pone.0113049

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molec Biol Evol 24:1596–1599. doi:10.1093/molbev/msm092

Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, Tragoonrung S (2010) The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17:11–22. doi:10.1093/dnares/dsp025

Thornhill AH, Ho SYW, Külheim C, Crisp MD (2015) Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. Molec Phylogen Evol 93:29–43. doi:10.1016/j.ympev.2015.07.007

Wilson PG, O'Brien MM, Heslewood MM, Quinn CJ (2005) Relationships within Myrtaceae sensu lato based on a *mat*K phylogeny. Pl Syst Evol 251:3–19. doi:10.1007/s00606-004-0162-y

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058. doi:10.1073/pnas.84.24.9054

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. doi:10.1093/bioinformatics/bth352

Genetics and Molecular Biology, 40, 4, 871-876 (2017)

DOI: http://dx.doi.org/10.1590/1678-4685-GMB-2017-0096

Genome Insight

Complete sequence and comparative analysis of the chloroplast genome of *Plinia trunciflora*

Maria Eguiluz[1], Priscila Mary Yuyama[2], Frank Guzman[2], Nureyev Ferreira Rodrigues[1] and Rogerio Margis[1,2]

[1]*Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.*
[2]*Departamento de Biofísica, Centro de Biotecnologia, Laboratório de Genomas e Populações de Plantas, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.*

Abstract

Plinia trunciflora is a Brazilian native fruit tree from the Myrtaceae family, also known as jaboticaba. This species has great potential by its fruit production. Due to the high content of essential oils in their leaves and of anthocyanins in the fruits, there is also an increasing interest by the pharmaceutical industry. Nevertheless, there are few studies focusing on its molecular biology and genetic characterization. We herein report the complete chloroplast (cp) genome of *P. trunciflora* using high-throughput sequencing and compare it to other previously sequenced Myrtaceae genomes. The cp genome of *P. trunciflora* is 159,512 bp in size, comprising inverted repeats of 26,414 bp and single-copy regions of 88,097 bp (LSC) and 18,587 bp (SSC). The genome contains 111 single-copy genes (77 protein-coding, 30 tRNA and four rRNA genes). Phylogenetic analysis using 57 cp protein-coding genes demonstrated that *P. trunciflora, Eugenia uniflora* and *Acca sellowiana* form a cluster with closer relationship to *Syzygium cumini* than with *Eucalyptus*. The complete cp sequence reported here can be used in evolutionary and population genetics studies, contributing to resolve the complex taxonomy of this species and fill the gap in genetic characterization.

Keywords: Jaboticaba, Myrtaceae, chloroplast genome, next-generation sequencing.

Received: April 18, 2017; Accepted: July 13, 2017.

Plinia trunciflora (O.Berg) Kausel, synonym *Myrciaria trunciflora* O.Berg, is a native Brazilian tree that belongs to the Myrtaceae family and is widely distributed in the southern and southeastern areas of Brazil (Sobral *et al.*, 2012). Among all identified *Plinia* sp. species, *P. cauliflora* (DC.) Berg (synonym *M. cauliflora* (Mart.) O.Berg), *P. jaboticaba* (Vell.) Berg (synonym *M. jaboticaba* O.Berg) and *P. trunciflora* are endemic to Brazil. All of these species produce a similar grape-like edible fruit, known as jaboticaba, which presents a sweet jelly-like white pulp covered by a purple peel. Jaboticaba (*P. trunciflora*) has attracted attention because of its significant levels of phenolic compounds associated with health benefits, such as antidepressant and antioxidant effects and the prevention of neurodegenerative diseases and diabetes (Stasi and Hiruma-Lima, 2002; Sacchet *et al.*, 2015). These benefits have largely been attributed to the capacity of these compounds to prevent or reduce oxidative stress. Additionally, jaboticaba (*P. trunciflora*) is largely consumed fresh or used to make jellies, juices, wines, spirits and vinegar (Balerdi *et al.*, 2006).

Send correspondence to Rogerio Margis. Departamento de Biofísica, Centro de Biotecnologia, Laboratório de Genomas e Populações de Plantas, Universidade Federal do Rio Grande do Sul (UFRGS), Avenida Bento Gonçalves 9500, Prédio 43432, Sala 206, Porto Alegre, RS, CEP 91501-970 Brazil. E-mail: rogerio.margis@ufrgs.br.

Despite the nutritional and productive recognized importance of this species, the taxonomic classification is still controversial. This is mostly so because it is based on morphological evaluation of the trees, fruits and seeds, regarding physical, chemical, physicochemical, and germinal characters that have shown the existence of variability (Guedes *et al.*, 2014). Therefore, molecular studies are needed to better clarify the phylogenetic relationships among the species from this genus.

The chloroplast (cp) genome is a circular molecule of double-stranded DNA that consists of four distinct regions, a large and a small single copy region (LSC and SSC, respectively) separated by two inverted repeat regions (IRa and IRb). Despite the high degree of conservation in its structure, gene content and organization, the presence of mutations, duplications and rearrangements of genes make it an attractive option for phylogenetic studies (Costa *et al.*, 2016). In the case of Myrtaceae, there are only few phylogenetic and evolutionary studies based on cp genes (Craven and Biffin 2005; Payn *et al.*, 2007; Biffin *et al.*, 2010; Bayly

et al., 2013; Eguiluz *et al.*, 2017; Machado *et al.*, 2017), and there are even less that include the *Plinia* genus (Vasconcelos *et al.*, 2017).

In this study, young leaves from a *Plinia trunciflora* tree harvested in Gravataí, RS, Brazil (latitude (S): 29°51'52"; longitude (W): 50°53'53") were used to extract total DNA by the CTAB method (Doyle and Doyle, 1990). DNA quality was evaluated by electrophoresis in a 1% agarose gel, and DNA quantity was determined using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). One genomic paired-end library of 100 nt length was generated by Fasteris SA (Plan-les-Ouates, Switzerland) using an Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA, USA). The paired-end sequence reads were filtered against 42 Myrtaceae cp genomes (Table S1) using BWA software with two mismatches allowed (Li and Durbin, 2009). The obtained reads were assembled *de novo* with ABySS software (Simpson *et al.*, 2009). The cp genome scaffolds were orientated using cp genome sequences of *Eucalyptus globulus*, *Eucalyptus grandis* and *Eugenia uniflora* L. using BLASTN (Camacho *et al.*, 2009). A gap region was filled in by Sanger sequencing using primers F: 5' GGGTTATCCTGCACTTGGAA and R: 3' TGCTGTCGAAGCTCCATCTA. Genes were annotated using DOGMA (Wyman *et al.*, 2004) and BLAST homology searches. tRNAs (transfer RNA) were predicted using tRNAscan-SE program (Schattner *et al.*, 2005) and confirmed by comparison with the appropriate homologs in *E. globulus*. The circular cp genome map was drawn using OGDRAW online program (Lohse *et al.*, 2007). For the phylogenetic analysis, a set of 57 cp protein-coding sequences (Table S2) from 56 species belonging to Malvids (Eurosids II) (Table S3) were used with *Vitis vinifera* serving as outgroup. Nucleotide sequences were aligned using MUSCLE available in MEGA version 6.0 (Tamura *et al.*, 2013), and a Bayesian tree was generated using MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003) with 5,000,000 generations sampled every 100 generations and discarding the first 25% of trees as burn-in, with posterior probability (PP) values for each node. The GTR+I+G nucleotide substitution model determined by MODELTEST version 3.7 (Posada and Crandall, 1998) was used. The phylogenetic tree was rooted and visualized using FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

A total of 148,824,244 raw Illumina paired-end reads from the *P. truncliflora* nuclear genome were filtered against 42 Myrtaceae cp genomes. The 8,912,157 obtained reads were *de novo* assembled into non-redundant contigs and singletons covering about 99% of the genome (minimum coverage=144 reads, maximum coverage=18,789 reads). Two final large scaffolds were obtained and joined into a cp circular genome using Sanger sequencing. The complete cp genome of *P. trunciflora* is 159,512 bp in size and was submitted to GenBank (accession number: KU318111). The size is similar to that of other Myrtaceae species (Eguiluz *et al.*, 2017; Machado *et al.*, 2017). The cp genome included an LSC region of 88,097 bp, an SSC region of 18,587 bp and a pair of inverted repeats (IRa and IRb) of 26,414 bp each (Figure 1). Coding regions comprise 47.2%, 13.3% correspond to rRNAs and tRNAs, and 39.5% of the genome comprises non-coding regions, including introns, pseudogenes and intergenic spacers (Table 1). In general, all genomic features showed similarity in structure and gene abundance with other Myrtaceae species (Bayly *et al.*, 2013; Eguiluz *et al.*, 2017; Machado *et al.*, 2017). The genome contained 131 genes in total, which includes 111 single-copy genes corresponding to 77 protein-coding genes, 30 transfer RNA (tRNA) genes and four ribosomal genes (rRNA) (Figure 1, Table 1). The *ycf1*, *ycf2* and *ycf15* sequences were annotated as pseudogenes based on the presence of many stop codons in their coding sequences and by comparison with sequences of *E. globulus* and *S. cumini*. Of the 131 genes in *P. trunciflora*, seven of the tRNAs genes and all four rRNA genes occurred within the IR regions and consequently were duplicated (Table 1). The cp genome has 20 intron-containing genes: 12 protein coding genes and six tRNA genes which contain one intron, and the *clpP* and *ycf3* genes that contain two introns each. The *rps12* gene is a trans-spliced gene with the 5'end located in the LSC region and the duplicated 3'end in the IR

Table 1 - Summary of the *Plinia trunciflora* chloroplast genome characteristics.

Feature	*Plinia trunciflora*
Total cpDNA size	159,512 bp
LSC size (bp)	88,097 bp
SSC size (bp)	18,586 bp
IR size (bp)	26,414 bp
Protein coding regions (%)	60.48%
rRNA and tRNA (%)	13.3%
Introns size (% total)	10.65%
Intergenic sequences and pseudogenes size (%)	28.9%

Feature	*Plinia trunciflora*
Number of genes	131 genes
Number of different protein coding genes	77
Number of different tRNA genes	30
Number of different rRNA genes	4
Number of different duplicated genes	16
Pseudogenes	3
GC content (%)	37%

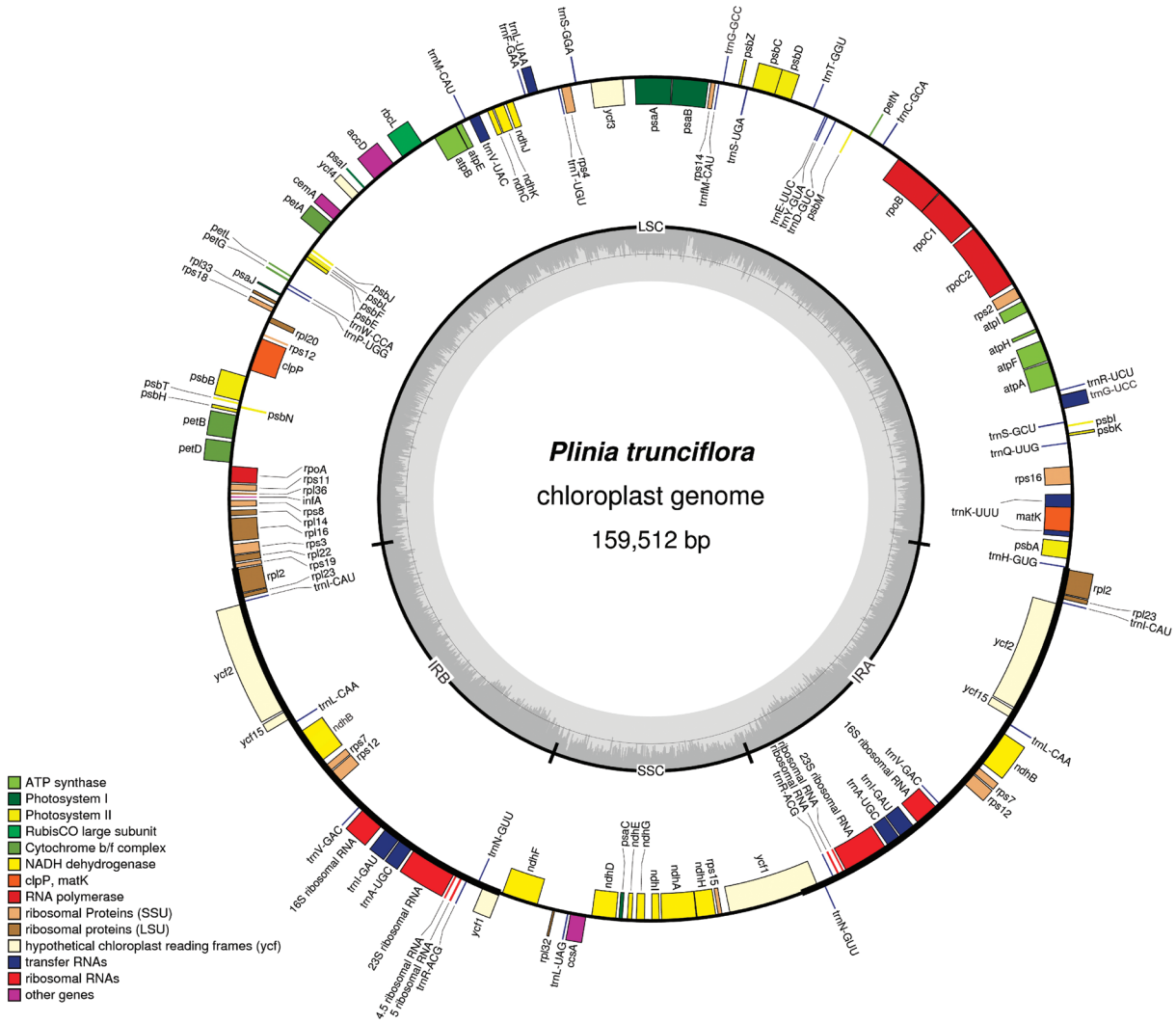


Figure 1 - Gene map of the *Plinia trunciflora* chloroplast genome. The structure of the cp genome consists of one large and small single copy (LSC and SSC, respectively) and a pair of inverted repeats (IRa and IRb). Genes drawn inside the circle are transcribed counterclockwise and those outsides are clockwise. Genes belonging to different functional groups are indicated by different tonalities. The darker gray in the inner circle corresponds to GC content, while the lighter gray corresponds to AT content.

regions. The *trnK-UUU* has 2,529 bp, with the largest intron encompassing also the *matK* gene.

The whole cp genome analysis revealed that the cp genomes of *P. trunciflora* and *E. uniflora* are shorter in comparison to other Myrtaceae, such as *E. globulus*, *E. grandis, E. uniflora* and *S. cumini*, (Figure 2). Despite its size, the total length of introns in *P. trunciflora* (16,972 pb) is the largest in Myrtaceae, e.g. *S. cumini* presents 14,469 bp and the same is observed in *E. globulus* and *E. grandis*. The size of the intergenic spacer located between the IRa/LSC border and the first gene of LSC in *P. trunciflora* is more similar to *Eucalyptus* species than its closer species *E. uniflora* (Figure 2). The comparison of the *ndhK* gene of *P. trunciflora*, with 678 bp, indicated a smaller gene size than that in other plants, such as *E. uniflora* (858 pb), *S. cumini* (855 bp), *E. globulus* (855 bp) and *E. grandis* (853 bp). The same size (678 bp) for this gene is found in *Arabidopsis thaliana*. The effective size of the coding sequence is confirmed by the presence of a thymine in position 53,811 bp in the cp genome from *P. trunciflora* that creates a stop codon and makes this gene shorter than in other Myrtaceae.

Our phylogeny includes the sister relationship of the orders Brassicales, Malvales and Sapindales and the orders Geraniales and Myrtales. All these results agree with previous studies based on multiple genes or complete cp genomes (Ruhfel *et al.*, 2014). By analyzing the Myrtaceae family clade we showed that *P. trunciflora*, *E. uniflora* and *Acca sellowiana* form a single cluster of Neotropical Myrtaceae, and that this clade has a shorter genetic distance with *S. cumini* than to the Australian Myrtaceae clade (Figure 3). Additionally, our analysis corroborates that *Corymbia gummifera* is paraphyletic in respect to *Angophora*. A previous phylogenetic analysis using certain cp

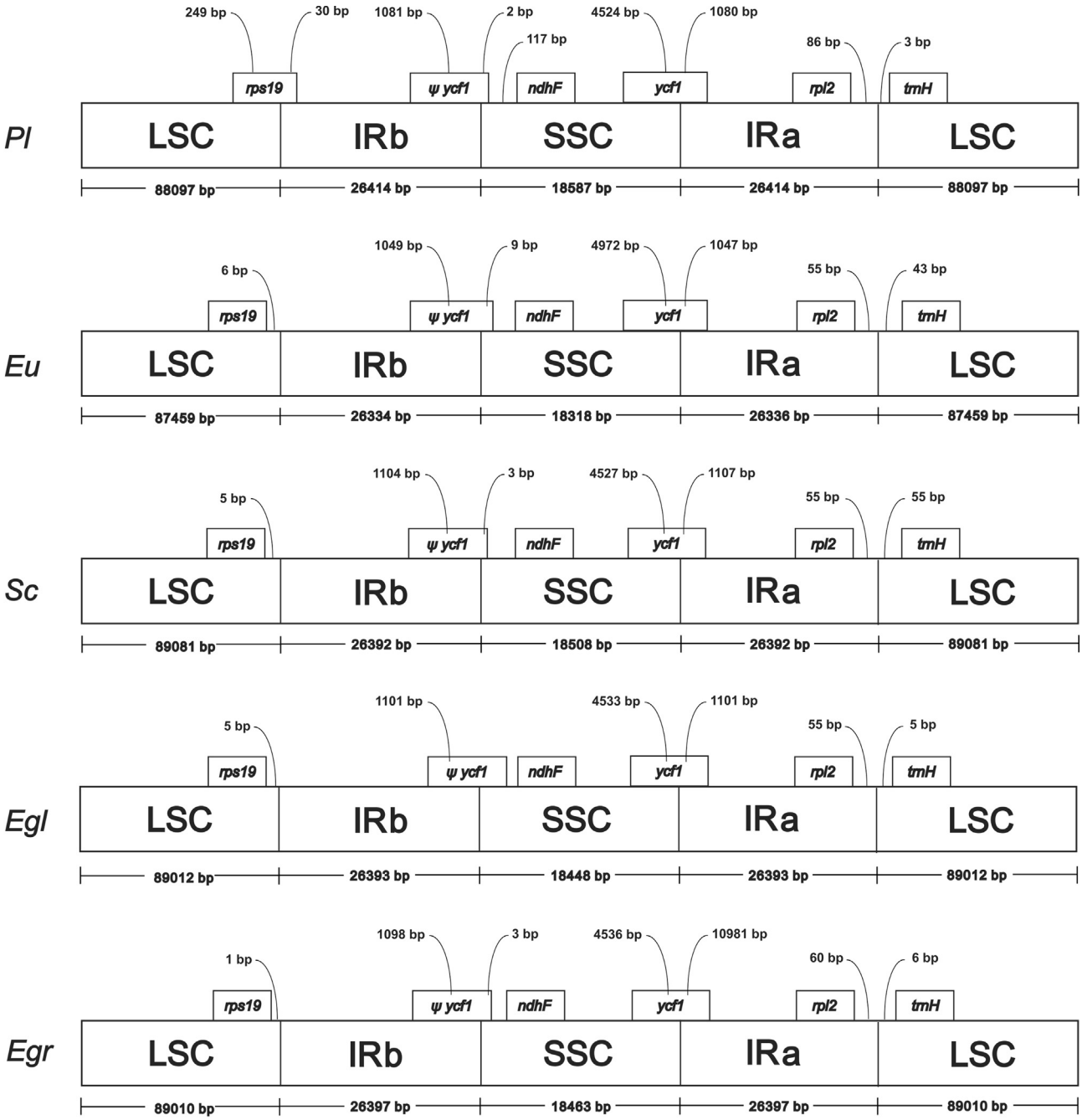


Figure 2 - Comparison of the borders of LSC, SSC and IR regions among five chloroplast genomes. Boxes above the main line indicate the predicted genes, while pseudogenes at the borders are shown by Ψ. Variation in *rps19* gene length is displayed at the IRb/LSC borders of *Plinia trunciflora*, *Eugenia uniflora*, *Syzygium cumini*, *Eucalyptus globulus* and *Eucalyptus grandis*, but only in *P. trunciflora*, this gene is located at IRb and LSC regions. This figure is not drawn to scale.

genes (ITS, *matK* and *ndhF*) of Myrtaceae species showed that *Eucalyptus*, *Syzygium*, *Eugenia* and *Myrciaria* (synonym of *Plinia*) form a distinct clade that is consistent with characteristics of the pollen (Thornhill *et al.*, 2012). As can be observed in the Bayesian tree (Figure 3), *Plinia* could be paraphyletic in relation to *Eugenia* and *Acca*, in agreement with the embryo morphology and studies using cp regions that placed *Plinia*, *Myrciaria* and *Siphoneugena* as the emerging "Plinia group" (Lucas *et al.*, 2007). Taxon sampling and phylogenetic methodology could affect the different results. Therefore, additional complete cp genome sequences will help in the comprehension of the relationship among Myrtaceae species.

The *Plinia trunciflora* genome represents the first complete cp genome sequence for the genus *Plinia* and shows a set of features that could be further explored for

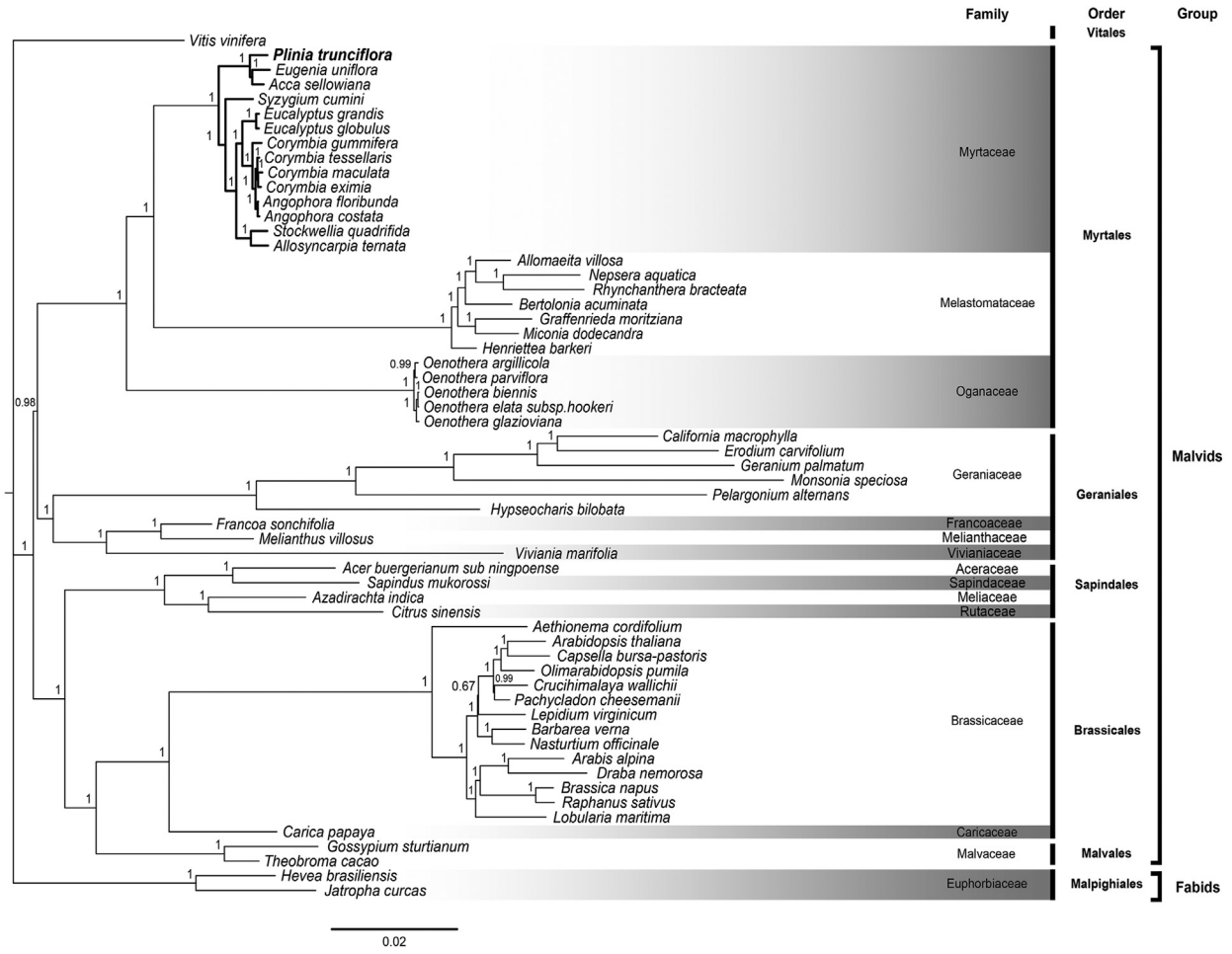


Figure 3 - Phylogenetic tree of Eurosids II based on 57 cp protein-coding genes generated by Bayesian method from 56 species. Bold branches indicate the Myrtaceae species. Numbers above each node are posterior probability values. Family, order and clade are also indicated. *Vitis vinifera* was considered as outgroup.

population and phylogenetic studies within this group. Moreover, these data increase the genetic and genomic resources available in Myrtaceae by adding a new strategy of organelle genome assembly.

Acknowledgments

This study was carried out with financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do Rio Grande do Sul (FAPERGS).

References

Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, Bossinger G, Merchant A, Udovicic F, Woodrow IE, *et al.* (2013) Chloroplast genome analysis of Australian eucalypts - Eucalyptus, Corymbia, Angophora, Allosyncarpia and Stockwellia (Myrtaceae). Mol Phylogenet Evol 69:704-716.

Balerdi CF, Rafie R and Crane J (2006) Jaboticaba (*Myrciaria cauliflora*, Berg.) a delicious fruit with an excellent market potential. Proc Florida State Hortic Soc 119:66-68.

Biffin E, Lucas EJ, Craven L, Da Costa IR, Harrington MG and Crisp MD (2010) Evolution of exceptional species richness among lineages of fleshy-fruited Myrtaceae. Ann Bot 106:79-93.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10:421.

Costa JF, Lin SM, Macaya EC, Fernández-García C and Verbruggenet H (2016) Chloroplast genomes as a tool to resolve red algal phylogenies: A case study in the Nemaliales. BMC Evol Biol 16:205.

Craven LA and Biffin E (2005) *Anetholea anisata* transferred to, and two new Australian taxa of *Syzygium* (Myrtaceae). Blumea 50:157-162.

Doyle JJ and Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13-15.

Eguiluz M, Rodrigues FN, Guzman F, Yuyama P and Margis R (2017) The chloroplast genome sequence from *Eugenia uniflora*, a Myrtaceae from Neotropics. Plant Syst Evol doi: 10.1007/s00606-017-1431-x.

Guedes MNS, Rufini JCM, Azevedo AM and Pinto NAVD (2014) Fruit quality of jabuticaba progenies cultivated in a tropical climate of altitude. Fruits 69:449-458.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Lohse M, Drechsel O and Bock R (2007) Organellar Genome DRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52:267-274.

Lucas EJ, Harris SA, Mazine FF, Belsham SR, Lughadha EMN, Telford A, Gasson PE and Chase MW (2007) Suprageneric phylogenetics of Myrteae, the generically richest tribe in Myrtaceae (Myrtales). Taxon 56:1105-1128.

Machado LO, Vieira LD, Stefenon VM, Pedrosa OF, De Souza EM, Guerra MP and Nodari RO (2017) Phylogenomic relationship of feijoa (*Acca sellowiana* (O.Berg) Burret) with other Myrtaceae based on complete chloroplast genome sequences. Genetica 145:1-12.

Payn KG, Dvorak WS and Myburg AA (2007) Chloroplast DNA phylogeography reveals the island colonisation route of *Eucalyptus urophylla* (Myrtaceae). Aust J Bot 55:673-683.

Posada D and Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. Bioinformatics 14:817-818.

Ronquist F and Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE and Burleigh JG (2014) From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol 14:23.

Sacchet C, Mocelin R, Sachett A, Bevilaqua F, Chitolina R, Kuhn F, Boligon AA, Athayde ML, Roman Junior WA, Rosemberg DB, *et al.* (2015) Antidepressant-like and antioxidant effects of *Plinia trunciflora* in mice. Evid Based Complement Alternat Med 2015:601503.

Stasi LC and Hiruma-Lima CA (2002) Myrtales medicinais. In: Stasi LC and Hiruma-Lima CA (eds) Plantas Medicinais na Amazônia e na Mata Atlântica. 2nd edition. Editora UNESP, São Paulo, pp 321-330.

Schattner P, Brooks AN and Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686-W689.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM and Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19:1117-1123.

Tamura K, Stecher G, Peterson D, Filipski A and Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30:2725-2729.

Thornhill AH, Hope GS, Craven LA and Crisp MD (2012) Pollen morphology of the Myrtaceae. Part 4: Tribes Kanieae, Myrteae and Tristanieae. Aust J Bot 60:260-289.

Wyman SK, Jansen RK and Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252-3255.

Vasconcelos NCT, Proença EBC, Ahmad B, Aguilar SD, Aguilar R, Amorim SB, Campbell K, Costa RI, De-Carvalho SP, Faria EQJ, *et al.* (2017) Myrteae phylogeny, calibration, biogeography and diversification patterns: Increased understanding in the most species rich tribe of Myrtaceae. Mol Phylogenet Evol 109:113-137.

Internet Resources

Sobral M, Proença C, Souza M, Mazine F and Lucas E (2012) Myrtaceae in lista de espécies da flora do Brasil. Jardim Botânico do Rio de Janeiro [online], http://floradobrasil.jbrj.gov.br (accessed 16 September 2015).

Supplementary material

The following online material is available for this article:

Table S1 - List of 42 Myrtaceae chloroplast genomes used in chloroplast genome assembling of *Plinia trunciflora.*

Table S2 - List of 57 chloroplast protein coding genes used in the phylogenetic analysis.

Table S3 - List of 56 plastome sequences of Rosids included in the Bayesian phylogenetic analysis.

Associate Editor: Guilherme Corrêa de Oliveira

Ann Microbiol (2017) 67:479–490
DOI 10.1007/s13213-017-1277-z

ORIGINAL ARTICLE

Comparative transcriptomic analysis of *Listeria monocytogenes* reveals upregulation of stress genes and downregulation of virulence genes in response to essential oil extracted from *Baccharis psiadioides*

Luiza Pieta[1] · Frank Lino Guzman Escudero[2] · Ana Paula Jacobus[3] · Kamila Patikowski Cheiran[1] · Jeferson Gross[3] · Maria Lisseth Eguiluz Moya[4] · Geraldo Luiz Gonçalves Soares[5] · Rogério Margis[2,4] · Ana Paula Guedes Frazzon[6] · Jeverson Frazzon[1]

Received: 3 March 2017 / Accepted: 8 May 2017 / Published online: 28 May 2017

Abstract *Listeria monocytogenes* is a pathogenic microorganism in humans and is frequently transmitted by food. Methods to control the presence of Listeria in foods are necessary. In the present study, transcriptomics of *L. monocytogenes* grown in the presence of essential oil extracted from *Baccharis psiadioides* were studied by RNA sequencing and reverse transcription quantitative polymerase chain reaction (RT-qPCR) experiments. The results obtained indicate that essential oil of *B. psiadioides* has potential bacteriostatic activity at the concentration tested, affecting *Listeria* cells functioning and development. Responses of the microorganism included upregulation of stress genes and downregulation of virulence genes, such as *actA*, *hly* and *prfA*, indicating a decrease in virulence and in the capacity of the microorganism to cause infection. Thus, the results presented here allow us to conclude that *B. psiadioides* essential oil may be an alternative means of controlling microorganisms proliferating in foods.

Keywords Bacteriostasis · Essential oil · *Listeria monocytogenes* · Virulence · RNA sequencing · RT-qPCR

Introduction

Among studies involving food safety, *Listeria monocytogenes* stands out because of its high pathogenicity, mainly related to immunocompromised individuals, such as the elderly and neonates, and the high risk of its transplacental transmission in pregnant women (Allerberger and Wagner 2010; Girard et al. 2014). The microorganism has the ability to survive and proliferate at refrigeration temperatures, which is a major problem related to food production that extensively uses the cold chain in the processing and storage of products (Farber and Peterkin 1991). Moreover, increased transcription of several *L. monocytogenes* genes involved in virulence and stress

Electronic supplementary material The online version of this article (doi:10.1007/s13213-017-1277-z) contains supplementary material, which is available to authorized users.

✉ Jeverson Frazzon
jeverson.frazzon@ufrgs.br

[1] Postgraduate Program in Food Science and Technology, Food Science and Technology Institute (ICTA), Federal University of Rio Grande do Sul (UFRGS), Bento Gonçalves Ave. 9500 / Building, 43212 Porto Alegre, Rio Grande do Sul (RS), Brazil

[2] Postgraduate Program in Cellular and Molecular Biology, Biotechnology Center (CBiot), Federal University of Rio Grande do Sul (UFRGS), Bento Gonçalves Ave. 9500 / Building, Porto Alegre, RS 43431, Brazil

[3] Institute for Research in Bioenergy, São Paulo State University (UNESP), 10th St. 2527, Rio Claro, São Paulo (SP), Brazil

[4] Postgraduate Program in Genetics and Molecular Biology, Federal University of Rio Grande do Sul (UFRGS), Bento Gonçalves Ave. 9500 / Building 43323M, Porto Alegre, RS, Brazil

[5] Department of Botany, Biosciences Institute , Federal University of Rio Grande do Sul (UFRGS), Bento Gonçalves Ave. 9500 / Building, 43433 Porto Alegre, RS, Brazil

[6] Department of Microbiology, Immunology and Parasitology, Basic Health Sciences Institute (ICBS), Federal University of Rio Grande do Sul (UFRGS), Sarmento Leite St, Porto Alegre, RS 500, Brazil

responses has already been demonstrated at 7 °C compared to 37 °C (Pieta et al. 2014). Among the 13 described serotypes of *L. monocytogenes*, 1/2a, 1/2b and 4b are responsible for 95% of human infections, called listeriosis (Montero et al. 2015). Historically, serotype 4b has caused the greatest proportion of listeriosis outbreaks and the largest number of cases per outbreak in the United States (Cartwright et al. 2013).

Essential oils (EO) are secondary metabolites produced by several plants, and can function as antimicrobials, antivirals, antimycotics, antipsoriatics, insecticides and in cancer treatments (Cowan 1999; Edris 2007; Reichling et al. 2009). The EO present in the *Asteraceae* plant family, with emphasis on *Baccharis psiadioides* (Less.) Joch. Müller (= *Heterothalamus psiadioides* Less.) (Giuliano and Freire 2011), has important anti-inflammatory properties (Fabri et al. 2011) and the ability to inhibit the growth of antibiotic resistant microorganisms, also reducing biofilm formation in abiotic surfaces (Negreiros et al. 2016). Natural compounds present in the essential oil of *B. psiadioides* (EO*Bp*) are classified as terpenes, and can be divided into two fractions: (1) monoterpenes with a significant percentage composed of β-pinene; and (2) sesquiterpenes with Ar-curcumene as the major component.

Transcriptomic, proteomic, genetic and physiological analyses can identify *L. monocytogenes* molecular stress adaptation responses, by global expression changes in a large number of the cellular components (Soni et al. 2011). In addition to EO, nisin—a bacteriocin produced by several lactic acid bacteria (Delves-Broughton 1990)—presents antimicrobial potential against food pathogens. Proteomic analyses of *L. monocytogenes* cells treated with a sub-lethal concentration of nisin displayed an overexpression of proteins related to oxidative stress and production of cell membrane lipids (Miyamoto et al. 2015). Experiments carried out with the Gram-positive pathogenic bacterium *Staphylococcus aureus*, showed transcriptional alterations induced by tea tree oil produced as a steam distillate of *Melaleuca alternifolia*, which has broad-spectrum antibacterial activity, including altered regulation of genes involved in heat shock and cell wall metabolism (Cuaron et al. 2013). Furthermore, the mechanism of biofilm inhibition and virulence attenuation in enterohemorrhagic *Escherichia coli* O157:H7 (EHEC) treated with eugenol and eugenol-rich oil was shown through transcriptional and phenotypic assays (Kim et al. 2016).

The use of natural compounds with antimicrobial potential represents an alternative means to combat pathogen growth; therefore, the present work aimed to analyze the differential transcriptome profile of *L. monocytogenes* grown in the presence of EO*Bp* using RNA sequencing (RNA-Seq) and reverse transcription quantitative polymerase chain reaction (RT-qPCR).

Materials and methods

Bacterial strain

The *L. monocytogenes* 55 (*Lm*55) strain was isolated from cheese by the National Agricultural Laboratory of Rio Grande do Sul State (LANAGRO/RS) of the Ministry of Agriculture, Livestock and Food Supply (MAPA/Brazil), and serotyped at the Oswaldo Cruz Institute (State of Rio de Janeiro, RJ, Brazil) as serotype 1/2a (de Mello et al. 2008; Nes et al. 2010).

Characterization of EO of *B. psiadioides*

EO*Bp* was obtained from the Laboratory of Chemical Ecology and Chemotaxonomy [Department of Botany, Federal University of Rio Grande do Sul (UFRGS)]. Leaves of *B. psiadiodes* were collected from populations located in Porto Alegre, RS, and subjected to drying at room temperature, with subsequent extraction of EO in a modified Clevenger apparatus (Gottleib and Taveira-Magalhães 1960). EO*Bp* was fractionated according to Kulisic et al. (2004) with some modifications, by column chromatography (40 cm in length; 2 cm diameter) with silica (21 g, 63–200 μm, 60° pore; Sigma-Aldrich, St. Louis, MO), using pentane and diethyl ether to obtain fractions containing only non-polar and polar hydrocarbons, respectively. Fractions obtained were analyzed using gas chromatography–mass spectrometry (GC-MS). For the experiments, the whole extract (both fractions) was used in *L. monocytogenes* cultures.

Experimental design, RNA sequencing and statistical analyses

The *Lm*55 strain was cultivated in tryptone soy broth (TSB; HiMedia, Mumbai, Maharashtra, India) at 37 °C under agitation. The MIC/2 of EO*Bp* (Negreiros et al. 2016) was added in the exponential growth phase, when the microorganism had reached an optical density ($OD_{600\ nm}$) between 0.3 and 0.4, measured with an ultraviolet/visible spectrophotometer (Ultrospec 3100 Pro; Amersham Biosciences, Little Chalfont, UK). After 20 min, growth was interrupted and cells were washed with 300 μL 1X TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0; reagents from Sigma-Aldrich) and resuspended in 100 μL 1X TE buffer. As control conditions, a parallel experiment was conducted without EO*Bp*. Total RNA samples from *Lm*55 were isolated using the TRIzol® Reagent kit (Thermo Fisher Scientific, Waltham, MA), and spectrophotometer readings (ratio $OD_{260\ nm}/OD_{280\ nm}$) comprised values between 1.8 and 2.0 for all samples. Experiments were performed in biological triplicates and experimental quadruplicates.

Total RNA samples were prepared using the TruSeq Stranded mRNA Sample Preparation—Low Sample (LS) protocol from the TruSeq Stranded mRNA Library Preparation Kit (Illumina, San Diego, CA), and a pool of libraries was prepared for subsequent sequencing according to the TruSeq Stranded mRNA Sample Preparation Guide (Illumina). Sequencing of the pooled libraries was performed on MiSeq Gene and Small Genome Sequencer equipment (Illumina) using the MiSeq Reagent kit v3 150 cycles (Illumina) according to the manufacturer's instructions. Finally, 600 μL [570 μL of the pooled libraries and 30 μL (5%) of PhiX control solution] was added to the cartridge for subsequent sequencing.

The presence of adapters and quality of reads produced by RNA-Seq were determined for each library using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Based on these data, the Trim Galore! software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to eliminate sequences of reads with a quality below 30, as well as the sequences of the Illumina adapters. The cleaned reads were then anchored with TopHat2 (Kim et al. 2013) to the reference genome of *Lm*55 (Pieta et al. 2015; deposited in GenBank under the accession no. LKHO00000000), and the fragments per kilobase million (FPKM) values for all genes were calculated using Cufflinks (Trapnell et al. 2012). The counting tables of the reads mapped to each gene were generated by the featureCounts module of Subread software (Liao et al. 2013) , for sequence alignment files generated by TopHat2. To perform statistical analyses for differential expression, the counting tables were analyzed in the R Bioconductor DESeq2 package v.1.12.3 (Love et al. 2014). For each treatment comparison, all genes with $\log_2$foldchange greater than 1 and less than −1 were considered differentially expressed.. The protein sequences of these two groups of genes were functionally annotated with Blast2GO (Conesa et al. 2005), and the functional categories were visualized with the WEGO program (Ye et al. 2006). Sequences of the proteins were compared to the UniRef Enriched KEGG Orthology (UEKO) database (Guedes et al. 2011) using local BlastX (Altschul et al. 1997). The BlastX results were processed in the MySQL software (Oracle, Cupertino, CA), and the KEGG Orthology (KO) codes obtained were viewed on the iPATH2 web server (Yamada et al. 2011).

Relative gene expression

From total RNA, complementary DNA (cDNA) synthesis, recommended by Bustin et al. (2009), was performed according to Pieta et al. (2014), and relative gene expression was determined using RT-qPCR. Primers were using the GenScript tool (https://www.genscript.com/tools/real-time-pcr-tagman-primer-design-tool) based on the genes that were differentially expressed and related to virulence, stress response and transcription factors of the microorganism. Genes chosen for analysis in the present study were *actA*, *agrA*, *crp*, *degU*, fri, *fur*, *hly*, *iscR*, *malR*, *prfA*, *sigB*, and *sod* (Table 1 and Table S1 for functions of the coded proteins).

For RT-qPCR experiments, a solution containing 0.01–0.1 μM of each primer; 25 μM dNTPs (Promega, Madison, WI); 1X reaction buffer; 3 mM $MgCl_2$; 1X SYBR Green (Bio-Rad, Hercules, CA); 0.25 U Platinum *Taq* DNA polymerase (Thermo Fisher Scientific); and ultrapure Milli-Q water to complete the final volume of 10 μL was prepared. Standard curves were constructed with four points in twofold dilutions starting from a 1:50 cDNA concentration for each of the study primers to verify reaction efficiency in RT-qPCR experiments, determined with the StepOne v. 2.3 software based on slopes of plots and crossing points (Cps) versus log input of cDNA. For amplification, StepOnePlus™ Real Time–PCR System (Thermo Fisher Scientific) and 96-wells polystyrene microplates (Axygen Scientific, Union City, CA) were used. PCR was conducted at 94 °C for 5 min; 40 cycles at 94 °C for 15 s, 60 °C for 10 s, 72 °C for 15 s and 60 °C for 35 s; and a final melting curve between 50 and 99 °C (Δ0.1 °C/s). All experiments were performed in biological triplicates and experimental quadruplicates. The total volume present in each well was 20 μL, consisting of 10 μL diluted cDNA (1:50) and 10 μL reaction solution, and in the case of the negative control, a total volume consisting of 20 μL reaction solution.

Housekeeping genes *gap*, *rpoB* and *16SrRNA* (Table S2) were tested as candidates for RT-qPCR data normalization using the NormFinder algorithm (Andersen et al. 2004) and geNorm v. 3.5 software (Vandesompele et al. 2002). Relative expression of the genes was calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen 2001), considering the efficiency (E) of RT-qPCR reactions for each of the primers in the calculation of relative expression ($E^{-\Delta\Delta Ct}$), and statistical analyses were performed using one-way analysis of variance (ANOVA), at a significance level of 5%, using Statistica software (Statsoft, Tulsa, OK). When there was a statistically significant difference ($P < 0.05$) between C_t (threshold cycle) values of the control and study conditions, the genes were considered to be more transcribed ($E^{-\Delta\Delta Ct} > 1$) or less transcribed ($E^{-\Delta\Delta Ct} < 1$) during growth in the presence of *Bp*EO.

Results and discussion

Determination of EOBp composition by GC-MS

Total EO*Bp* was used to perform our analysis and the EO*Bp* fractions obtained were divided into two groups: one fraction was composed predominantly of monoterpenes and the other predominantly of sesquiterpenes. Results of GC-MS indicated the presence of a complex mixture of terpenes in the two

Table 1 Sequences of primers used in the transcriptional analysis by RT-qPCR, with respective sizes of amplification fragments and annealing temperatures

Gene	Nucleotide sequence	Amplicon size (bp)	Annealing temperature (°C)
actA	5′ AGAAATCATCCGGGAAACAG 3′	147	58.98
	5′ CCTCTCCCGTTCAACTCTTC 3′		58.87
agrA	5′ CGGGTACTTGCCTGTATGAA 3′	149	58.65
	5′ TGAATAGTTGGCGCTGTCTC 3′		59.03
crp	5′ ATTCACAGTTTGCGAATGCT 3’	117	58.86
	5′ TTTGCAAATCAACATCACGA 3’		59.02
degU	5′ GGCGCGTATATTCATCCAC 3’	150	58.96
	5′ TACCTCGCACTCTCTATGCG 3’		59.20
fri	5′ GGCGAACAAATGGATGAAGT 3’	108	59.94
	5′ ATAAGGCGCTTCTTCTACGC 3’		58.77
fur	5′ TTTAGCGCCTTCTTGTCTCA 3’	114	58.80
	5′ GGCCTTGCAACCGTTTATAG 3’		59.61
hly	5′ AGCTCATTTCACATCGTCCA 3’	124	59.24
	5′ TGGTAAGTTCCGGTCATCAA 3’		58.97
iscR	5′ ATCGGACCTCTTCGTAATGC 3’	106	59.15
	5′ CGTATGATATCACCCGCAGT 3’		58.48
malR	5′ GAATCGTCTGGACCGGTAAT 3’	110	58.86
	5′ AACGTGAGCCAAGTCCTTCT 3’		58.94
prfA	5′ GGAAGCTTGGCTCTATTTGC 3’	145	59.07
	5′ ACAGCTGAGCTATGTGCGAT 3’		58.65
sigB	5′ TGGTGTCACGGAAGAAGAAG 3’	135	58.85
	5′ TCCGTACCACCAACAACATC 3’		59.27
sod	5′ CCACCATTTGGGCTAAGAAT 3’	94	58.90
	5′ GCGTTCCTGAAGATATTCGC 3’		59.81

fractions. The fraction composed predominantly of monoterpenes revealed the presence of 20 compounds (Table 2); monoterpenes represented 71.82% of this fraction, with β-pinene as the major compound (43.81%). Other compounds present in significant amounts were δ-3-carene (14.92%) and limonene (10.82%)—both monoterpenes. In relation to the fraction composed predominantly by sesquiterpenes, the presence of 14 compounds was verified (Table 3), where the sesquiterpenes represented 93.59% of this fraction, Ar-curcumene being the major compound (40.12%). In this fraction, other compounds were also found in significant concentrations, such as bicyclogermacrene (15.89%) and γ-muurolene (15.68%)—both sesquiterpenes.

Transcriptomic analysis

In total, 333 genes presented a $\log_2$foldchange > −1 (−2 fold change cut off), being considered downregulated in the T4 sample (untreated with EO*Bp*), and, consequently, upregulated in the O6 sample (treated with MIC/2 EO*Bp*); and 273 genes presented a $\log_2$foldchange > 1 (2 fold change cut off), which means they were upregulated in the T4 and downregulated in the O6 samples (Table S3).

Based on these data, functional categories were visualized with the WEGO program, and the results regarding the effect of EO*Bp* on differential genes expression in *Lm*55 strain are shown in Fig. 1 and Table 4 for the three categories listed: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

With regard to the BP group (Fig. 1a), several processes presented a greater number of upregulated genes, such as biological regulation; cell cycle; catabolic process; amino acid and nitrogen compound, carbohydrate, cofactor, lipid, organic acid and sulfur metabolism; and response to stress. According to Bich et al. (2016), “biological regulation is what allows an organism to handle the effects of a perturbation, modulating its own constitutive dynamics in response to particular changes in internal and external conditions”. As the results showed 12 upregulated genes and 4 downregulated genes in this category, indicating that EO*Bp* can affect homeostasis causing changes in *L. monocytogenes* cells function and development. In support of this statement, growth in the presence of EO*Bp* upregulated 22 genes and downregulated 5 genes related to stress response. In addition, several genes related to cofactor and sulfur metabolism were upregulated, and it should be noted that the iron-sulfur ([Fe-S]) clusters or cofactors (widely distributed in nature) are of great importance in several biological processes (Johnson et al. 2005).

Table 2 Chemical composition of *Baccharis psiadioides* essential oil (EO*Bp*) fraction composed predominantly by monoterpenes. The relative percentage of each component was obtained directly from the peak areas of the chromatogram, considering 100% the sum of all evaluated peaks

Component	IK cal[a]	IK tab[b]	Yield (%)
Monoterpenes			
α-pinene	930	939	0.59
β-pinene	978	979	43.81
Mircene	993	990	0.93
δ-3-carene	1012	1011	14.92
p-cymene	1024	1024	0.75
Limonene	1029	1029	10.82
Total			71.82
Sesquiterpenes			
β-elemene	1383	1390	1.65
β-caryophyllene	1407	1419	1.15
Aromadendrene	1426	1441	1.80
Dehydro-aromadendrene	1434	1462	2.36
Allo-aromadendrene	1446	1460	3.34
γ-gurjunene	1457	1477	1.06
γ-muurolene	1462	1479	1.09
Germacrene D	1466	1481	0.80
Ar-curcumene + β-selinene	1470	1480/1490	5.32
Valencene	1473	1496	0.87
α-selinene	1480	1498	4.28
α-muurolene	1485	1500	0.90
γ-cadinene	1496	1513	1.53
δ-cadinene	1506	1523	2.03
Total			28.18

[a] Calculated Kováts retention index

[b] Tabulated Kováts retention index

Table 3 Chemical composition of EO*Bp* fraction composed predominantly by sesquiterpenes. The relative percentage of each component was obtained directly from the peak areas of the chromatogram, considering 100% the sum of all evaluated peaks

Component	IK cal[a]	IK tab[b]	Yield (%)
Monoterpenes			
β-pinene	973	979	1.41
p-cymene	1023	1024	0.64
Limonene	1027	1029	3.14
(E)-β-ocimene	1047	1050	1.22
Total			6.41
Sesquiterpenes			
β-elemene	1383	1390	4.30
β-caryophyllene	1407	1419	1.12
α-humulene	1440	1454	6.56
Allo-aromadendrene	1446	1460	4.91
γ-muurolene	1468	1479	15.68
Ar-curcumene	1477	1480	40.12
Bicyclogermacrene	1485	1500	15.89
Germacrene A	1491	1509	2.07
γ-cadinene	1498	1513	1.30
δ-cadinene	1508	1523	1.64
Total			93.59

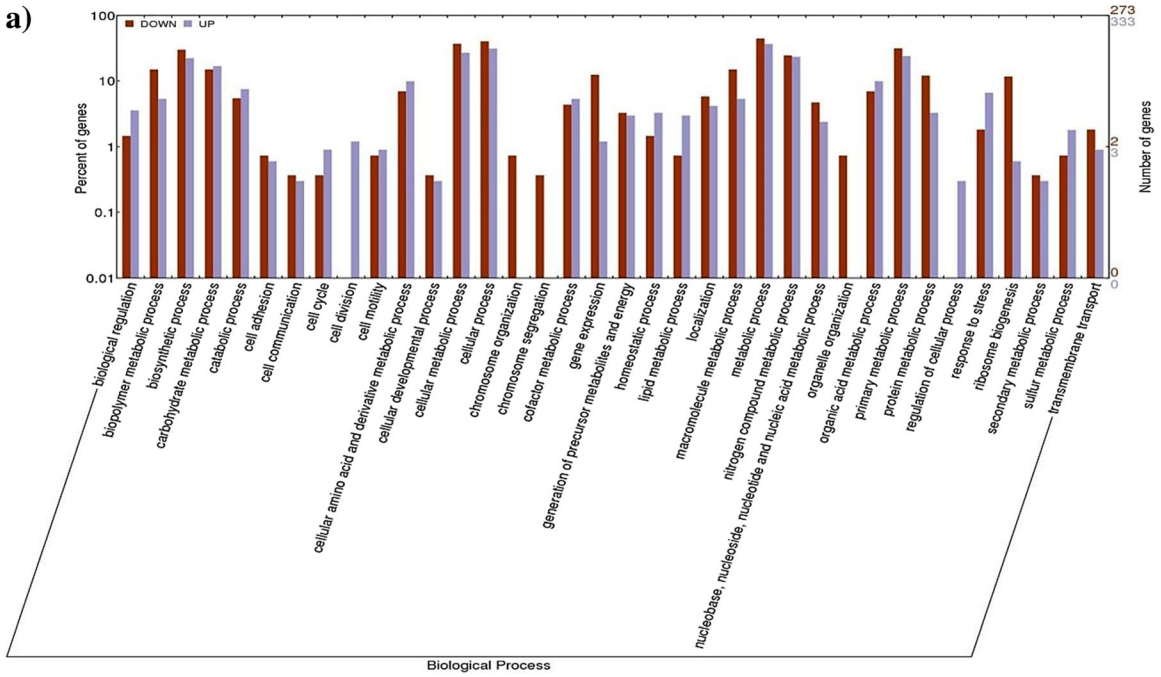
[a] Calculated Kováts retention index

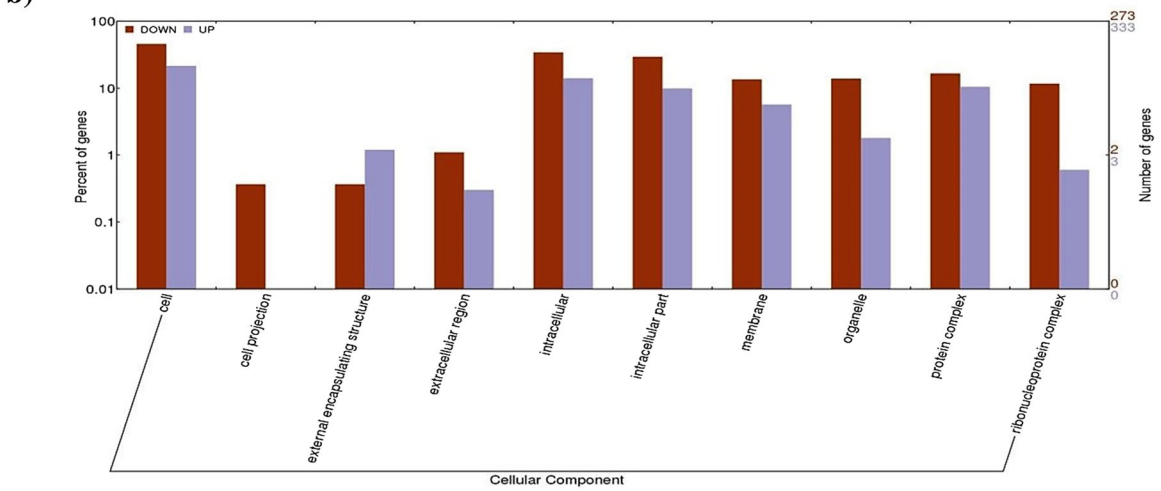
[b] Tabulated Kováts retention index

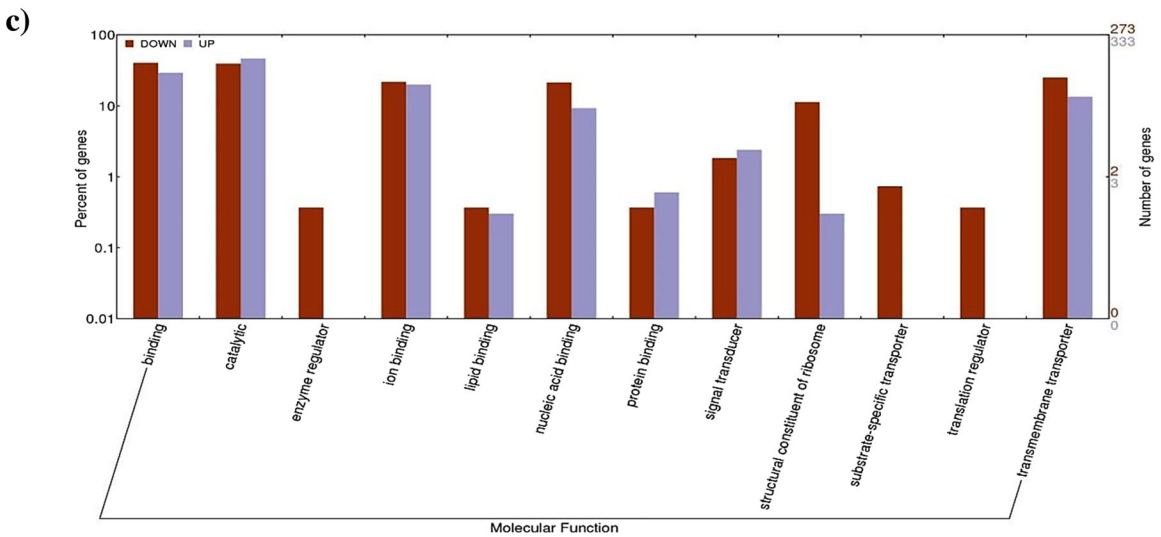
Carbohydrate and lipid metabolism indicate energy generation, and may be considered catabolic processes, which refer to the assimilation or processing of organic compounds to obtain energy. Positive regulation of genes involved in the metabolism of several compounds may be related to the EO composition, since EO are complex mixtures of volatile substances, usually lipophilic, whose components include terpene hydrocarbons, simple alcohols, aldehydes, ketones, phenols, esters and fixed organic acids (Simões and Spitzer 1999). Araújo et al. (2016) analyzed the effects of argentilactone, a constituent of the EO from *Hyptis ovalifolia*, on the transcriptional profile, cell wall and oxidative stress of *Paracoccidioides* spp., a dimorphic pathogenic fungus. Their results demonstrated that the upregulated genes were related to metabolism; cell rescue, defense and virulence; energy and cell cycle; and DNA processing. The downregulated genes were related to metabolism, transcription, protein fate, and cell cycling and DNA processing.

A larger number of downregulated genes related to BP were identified for categories such as biopolymers, macromolecules and protein metabolism; cell division; gene expression; ribosome biogenesis; and transmembrane transport. Biopolymer metabolism includes proteins, DNA and RNA production, and its downregulation may consequently affect ribosome biogenesis (32 downregulated versus two upregulated genes) and gene expression (34 downregulated versus four upregulated genes). The antimicrobial effect of EO may be responsible for downregulation of genes related to cell division, indicating the difficulty that the microorganism has, in the presence of the EO, to complete its binary fission and increase the microbial population.

All the categories related to CC (Fig. 1b) presented a larger number of downregulated genes, except for the external encapsulating structure. Some of those belonging to MF (Fig. 1c), such as structural constituent of ribosomes, translation regulators and transmembrane transporters, were also mostly downregulated. These data suggest an inverse correlation with the results for higher numbers of downregulated genes involved in BP, such as ribosome biogenesis, biopolymer (DNA, RNA, proteins) production, and transmembrane transport.


a)
DOWN
UP
Percent of genes
Number of genes
100
10
1
0.1
0.01
273
333
2
3
0
biological regulation
biopolymer metabolic process
biosynthetic process
carbohydrate metabolic process
catabolic process
cell adhesion
cell communication
cell cycle
cell division
cell motility
cellular amino acid and derivative metabolic process
cellular developmental process
cellular metabolic process
cellular process
chromosome organization
chromosome segregation
cofactor metabolic process
gene expression
generation of precursor metabolites and energy
homeostatic process
lipid metabolic process
localization
macromolecule metabolic process
metabolic process
nitrogen compound metabolic process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
organelle organization
organic acid metabolic process
primary metabolic process
protein metabolic process
regulation of cellular process
response to stress
ribosome biogenesis
secondary metabolic process
sulfur metabolic process
transmembrane transport
Biological Process


b)
DOWN
UP
Percent of genes
Number of genes
100
10
1
0.1
0.01
273
333
2
3
0
cell
cell projection
external encapsulating structure
extracellular region
intracellular
intracellular part
membrane
organelle
protein complex
ribonucleoprotein complex
Cellular Component


c)
DOWN
UP
Percent of genes
Number of genes
100
10
1
0.1
0.01
273
333
2
3
0
binding
catalytic
enzyme regulator
ion binding
lipid binding
nucleic acid binding
protein binding
signal transducer
structural constituent of ribosome
substrate-specific transporter
translation regulator
transmembrane transporter
Molecular Function

◀ **Fig. 1a–c** Transcriptomic analysis results. Differential expression of genes related to functional categories **a** biological process (BP), **b** cellular component (CC), and **c** molecular function (MF) of *Listeria monocytogenes* 55 grown in the presence of *Baccharis psiadioides* essential oil (EO*Bp*). Graphical representation generated using the WEGO program

Transcriptional analysis of virulence genes and stress response genes

First, to determine the reliability of the amplification data, the efficiency of the study primers was determined (Table S4), and the housekeeping genes *gap*, *rpoB* and *16SrRNA* were tested as candidates for RT-qPCR data normalization using the NormFinder algorithm and geNorm v. 3.5 software. Both programs indicated *rpoB* and *16SrRNA* as the most stable genes and recommendable for data analysis, while *gap* was demonstrated as the least stable gene (Fig. S1 and Table S5). Results of relative gene expression for *Lm*55 strain cultivated in the presence of EO*Bp* are shown in Fig. 2. The data shown here concur with the differential expression obtained with RNA-Seq, which allowed us to validate our experiments (Table S6).

Downregulation ($P < 0.05$) was observed in virulence genes, such as *prfA*, *fur*, *hly*, *actA* and *agrA*, in the presence of EO*Bp*. Previous research has already demonstrated the antimicrobial and antibiofilm potential of plant-extracted EO against several food-borne pathogens, such as *S. aureus*, *E. coli* and *L. monocytogenes* (Upadhyay et al. 2013; Lopez-Romero et al. 2015) and the relation between the EO concentration and its bactericidal and/or bacteriostatic effect against these bacteria (Burt 2004; Mazzarrino et al. 2015). In addition, the extracted of EO*Bp* showed a high concentration of β-pinene—a monoterpene that has been reported as one of the main chemicals responsible for the antimicrobial activity of several EOs.

Both PrfA and Fur are regulators involved in *L. monocytogenes* virulence and pathogenicity. PrfA controls the transcription of several virulence genes involved in the infection process, such as *actA*, which is responsible for the polymerization of actin tails, which propels the microorganism to neighboring cells, and the *hly* gene that codifies listeriolysin O (LLO), which is critical to survival of the microorganism in the phagocytes during the infection process (Xayarath and Freitag 2012). Thus, the significantly reduced transcription of *prfA* corroborates the reduced transcription of the *hly* gene. The *agr* system of *S. aureus*, widely conserved among Gram-positive bacteria, is involved in biofilm formation (Lyon and Novick 2004), and the AgrA-AgrC two-component system has been studied extensively because of its control of virulence factors (Novick 2000). In *L. monocytogenes*, as in *S. aureus*, *agrB*, *agrD*, *agrC* and *agrA* genes are organized in a unique operon, regulating microorganism adhesion to surfaces, fundamental for a proper biofilm formation, in addition to its involvement in the *Listeria* infection process in mammals (Riedel et al. 2009). An earlier in vivo study showed that the virulence of a Δ*agrA L. monocytogenes* strain was attenuated, demonstrating the role of the *agr* locus in the virulence of this microorganism, and its influence in the production of several secreted proteins, such as LLO (Autret et al. 2003).

Iron, an abundant element in nature, acts as a cofactor for several enzymes involved in microorganism metabolism, being required by almost all bacteria. However, iron concentrations above physiological levels can be toxic for microorganisms. A regulator of ferric iron uptake in many bacteria, Fur is involved with *L. monocytogenes* virulence and survival in the host (Rea et al. 2004). Mutations in the *fur* gene reduced microorganism pathogenicity in mice, indicating that disruption of intracellular iron homeostasis contributes to a lower ability of this pathogen to successfully establish infection (Newton et al. 2005; Olsen et al. 2005). In agreement with this, McLaughlin et al. (2012) demonstrated that deregulation of iron uptake through the elimination of Fur significantly impacts upon virulence potential in several pathogenic bacteria, including *L. monocytogenes*, as mutants in Fur-regulated loci resulted in a significant reduction in virulence potential relative to the wild-type. A recent study characterized the composition of an EO extracted from the leaf of *Rhaphiodon echinus* GC-MS experiments revealed the presence of monoterpenes, sesquiterpenes, and the metal chelation potential of this oil (Duarte et al. 2016). As the EO*Bp* constitutes by both monoterpenes and sesquiterpenes, this may explain the significantly decreased transcription of *fur*, which is downregulated under iron-limited conditions (Ledala et al. 2010).

While some genes associated with virulence were downregulated, genes correlated with stress response such as *degU*, *sigB*, *crp*, *fri*, *iscR*, *sod* and *malR* were upregulated in the presence of EO*Bp*. An upregulation gene example was a stress response transcription factor named sigma B (σ^B), which contributes to the microorganism's resistance to several conditions unsuitable to its development, such as acidic, osmotic and energy stresses (O'Byrne and Karatzas 2008).

DegU is a regulator of the expression of flagellar and chemotaxis genes in *L. monocytogenes*, involved in microorganism motility but not required for its virulence (Williams et al. 2005). Burke et al. (2014) demonstrated that *L. monocytogenes* uses different enzymes and regulators of gene expression, such as DegU, to resist the bactericidal activity of lysozymes, which degrade the bacterial cell wall, resulting in bacteriolysis. In addition, they suggested that DegU is one of the major regulators of lysozyme resistance in *L. monocytogenes*, a mechanism commonly found in other pathogens. Members of the Crp/Fnr transcription factor family have several related functions in microorganisms, such as regulation of virulence, metabolic pathways and stress response.

Table 4 Number of down and upregulated genes related to processes belonging to the functional categories studied [Biological Process (BP); Cellular Component (CC); Molecular Function MF)] in *Listeria monocytogenes* 55 grown in the presence of EO*Bp*

	Down	Up
BP		
Biological regulation	4	12
Biopolymer metabolic process	41	18
Biosynthetic process	82	74
Carbohydrate metabolic process	41	56
Catabolic process	15	25
Cell adhesion	2	2
Cell communication	1	1
Cell cycle	0	4
Cell division	45	35
Cell motility	2	3
Cellular amino acid and derivative metabolic process	19	33
Cellular developmental process	1	1
Cellular metabolic process	101	90
Cellular process	110	104
Chromosome organization	2	0
Chromosome segregation	1	0
Cofactor metabolic process	12	18
Gene expression	34	4
Generation of precursor metabolites and energy	9	10
Homeostatic process	4	11
Lipid metabolic process	2	10
Localization	16	14
Macromolecule metabolic process	41	18
Metabolic process	122	122
Nitrogen compound metabolic process	67	78
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	13	8
Organelle organization	2	0
Organic acid metabolic process	19	33
Primary metabolic process	86	80
Protein metabolic process	33	11
Regulation of cellular process	0	1
Response to stress	5	22
Ribosome biogenesis	32	2
Secondary metabolic process	1	1
Sulfur metabolic process	2	6
Transmembrane transport	5	3
Cellular component		
Cell	124	71
Cell projection	1	0
External encapsulating structure	1	4
Extracellular region	3	1
Intracellular	93	47
Intracellular part	80	33
Membrane	37	19
Organelle	38	6
Protein complex	45	35
Ribonucleoprotein complex	32	2
Molecular function		
Binding	110	97
Catalytic	107	154
Enzyme regulator	1	0
Ion binding	59	66
Lipid binding	1	1
Nucleic acid binding	58	31
Protein binding	1	2
Signal transducer	5	8
Structural constituent of ribossome	31	1
Substrate-specific transporter	2	0
Translation regulator	1	0
Transmembrane transporter	68	45

Crp, the cyclic AMP receptor protein, affects the metabolism of sugars or amino acids, transport processes, protein folding, as well as toxin production or pilus synthesis (Körner et al. 2003). In addition, the Crp family of transcription factors is

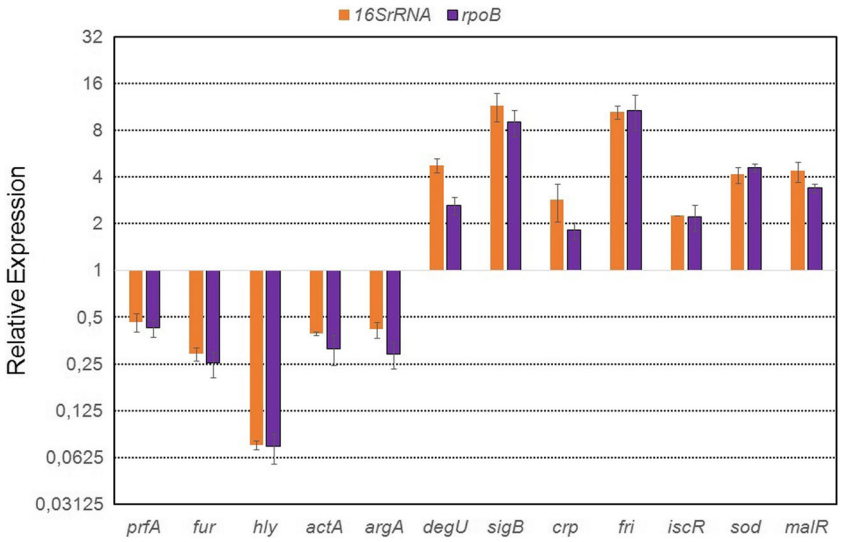


Fig. 2 Transcriptional analysis results. Relative expression of *actA*, *agrA*, *crp*, *degU*, *fri*, *fur*, *hly*, *iscR*, *malR*, *prfA*, *sigB* and *sod*, normalized with *rpoB* and *16SrRNA*, for *Listeria monocytogenes* 55 grown in the presence of EO*Bp*, and respective bars indicating the standard deviation values. All genes were statistically less or more transcribed ($P < 0.05$); graphical representation obtained with Microsoft Office Excel 2007

involved in various metabolic pathways in bacteria, acting in response to environmental changes. It has been shown that Crp acts as a transcription regulator in response to stresses in *Deinococcus radiodurans* (Yang et al. 2016). This Gram-positive bacterium is characterized by its efficient DNA repair ability and extreme stress resistance (Makarova et al. 2001) and generally considered to be an ideal model organism for studying bacterial resistance mechanisms under various stress conditions. This recent study demonstrated that the transcription levels of *crp* genes were increased to different extents when the bacteria were exposed to oxidizing agents. The Crp mutants were more susceptible to hydrogen peroxide (H_2O_2) than the wild-type strain, proving the important role of these proteins in stress resistance of *D. radiodurans*.

The *fri* gene encodes an iron-binding ferritin-like protein (Fri) that belongs to the Dps (DNA-binding proteins from starved cells) family of proteins (Haikarainen and Papageorgiou 2010). Ferritin is the most important iron reserve protein, found in all cells, especially in those involved in ferric compound synthesis, iron reserves and metabolism, which is required by several bacteria. It has been shown that the *fri* gene is repressed by Fur (Fiorini et al. 2008), being upregulated under several conditions, such as iron restriction, heat and cold shock (Hébraud and Guzzo 2000). The results obtained in the present study confirm this, since the *fur* gene was downregulated, and, consequently, the *fri* gene was upregulated in the presence of EO*Bp*. A recent study demonstrated that the cell-envelope stress response in *L. monocytogenes* is linked to the osmotic stress response, confirming the results obtained in the present work, because active terpenes compounds present in EO*Bp* act by binding the cell membrane of microorganisms (Milecka et al. 2015). Several studies suggest that Fri has a global impact on the *L. monocytogenes* regulatory network (Dussurget et al. 2005; Olsen et al. 2005), and this protein is also a mediator of beta-lactam tolerance and resistance to antibiotics such as cephalosporins (Krawczyk-Balska et al. 2012).

Iron is also necessary for cellular growth, development and survival, thus the [Fe-S] clusters—*isc*—are cofactors of enzymes involved in several biological processes related to respiration, DNA repair, carbon/nitrogen metabolism and regulation of gene expression (Py and Barras 2010). The *isc* operon encodes IscR, a [2Fe-2S] transcription factor that is involved in [Fe-S] cluster biogenesis, being a regulator responsible for governing various physiological processes during growth and stress responses (Mettert and Kiley 2014). IscR is widely conserved among proteobacteria (Rodionov et al. 2006); however, in Gram-positive bacteria, it is not well characterized. A relevant study performed by Santos et al. (2014) demonstrated that a gene from the unique Gram-positive dissimilatory metal-reducing bacterium *Thermincola potens*, which belongs to the *Firmicutes* phylum, the same as *Listeria* species, encodes a functional IscR homolog that is likely involved in the regulation of iron-sulfur cluster biogenesis.

Catalase (Kat) and superoxide dismutase (Sod) are the two major proteins implicated in protection against superoxides and reactive oxygen species (ROS) (Camejo et al. 2009), as the *sod* gene acts by dismutating the superoxide radical anion $O_2^{\bullet-}$ to H_2O_2, which is transformed into H_2O by the *kat* gene (Imlay 2003). Sod proteins can be classified into different types according to their metal cofactors, but only manganese-dependent superoxide dismutase (MnSod) is found in *L. monocytogenes* (Vasconcelos and Deneer 1994). In the present study, the *sod* gene was upregulated in the presence of EO*Bp*, in agreement with others studies related to the oxidative stress response. In addition to providing bacterial resistance against host-generated toxic oxygen species, *sod* gene induction has also been demonstrated during biofilm formation (Trémoulet et al. 2002), which is related to

oxidative stress in several bacteria as a response to changes in environmental conditions (Arce Miranda et al. 2011; Bitoun et al. 2011). As well as EO, ozone also has antimicrobial potential, being widely used in food processing due to its significant disinfection and ability to degrade rapidly. Both catalase and superoxide dismutase were found to protect pathogenic *L. monocytogenes* cells from ozone attack (Fisher et al. 2000).

Listeria species are widespread in the environment and soils, which are rich in complex carbohydrates like starch and its degradation products maltodextrins and maltose, requiring efficient uptake mechanisms for these compounds (Gopal et al. 2010). The maltose repressor protein (MalR) is a member of the LacI/GalR regulatory family, which is responsible for controlling a broad range of bacterial metabolic processes, from selective carbon source utilization to nucleotide synthesis and amino acid catabolism (Nguyen and Saier 1995; Swint-Kruse and Matthews 2009).

In conclusion, the use of natural compounds provides a new way for the scientific community to control the growth of microorganisms in food products. Results obtained in the present study on the antimicrobial effect of EO*Bp* on *Lm*55 isolated from dairy products (cheese), indicate a downregulation of virulence genes and upregulation of stress response genes, which results in destabilization of bacteria. *L. monocytogenes* is considered one of the pathogens with higher mortality rates involved in foodborne outbreaks, thus the possibility of reducing its pathogenicity becomes of great relevance for future research.

Acknowledgements We acknowledge the National Council for Scientific and Technological Development of Brazil (CNPq) (J. F. Grants #473181/2013-4 and #303603/2015-1).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Allerberger F, Wagner M (2010) Listeriosis: a resurgent foodborne infection. Clin Microbiol Infect 16:16–23

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Andersen CL, Jensen JL, Ørntoft TF (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res 64:5245–5250

Araújo FS, Coelho LM, Silva LC, Neto BRS, Parente-Rocha JA, Bailão AM, de Oliveira CMA, Fernandes GR, Hernández O, Ochoa JGM, Soares CMA, Pereira M (2016) Effects of argentilactone on the transcriptional profile, cell wall and oxidative stress of *Paracoccidioides* spp. PLoS Negl Trop Dis. doi:10.1371/journal.pntd.0004309

Arce Miranda JE, Sotomayor CE, Albesa I, Paraje MG (2011) Oxidative and nitrosative stress in *Staphylococcus aureus* biofilm. FEMS Microbiol Lett 315:23–29

Autret N, Raynaud C, Dubail I, Berche P, Charbit A (2003) Identification of the agr locus of *Listeria monocytogenes*: role in bacterial virulence. Infect Immun 71:4463–4471

Bich L, Mossio M, Ruiz-Mirazo K, Moreno A (2016) Biological regulation: controlling the system from within. Biol Philos 31:237–265

Bitoun JP, Nguyen AH, Fan Y, Burne RA, Wen ZT (2011) Transcriptional repressor Rex is involved in regulation of oxidative stress response and biofilm formation by *Streptococcus mutans*. FEMS Microbiol Lett 320:110–117

Burke TP, Loukitcheva A, Zemansky J, Wheeler R, Boneca IG, Portnoya DA (2014) *Listeria monocytogenes* is resistant to lysozyme through the regulation, not the acquisition, of cell wall-modifying enzymes. J Bacteriol 196:3756–3767

Burt S (2004) Essential oils: their antibacterial properties and potential applications in foods—a review. Int J Food Microbiol 94:223–253

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55:611–622

Camejo A, Buchrieser C, Couve E, Carvalho F, Reis O, Ferreira P, Sousa S, Cossart P, Cabanes D (2009) In vivo transcriptional profiling of *Listeria monocytogenes* and mutagenesis identify new virulence factors involved in infection. PLoS Pathog 5:e1000449. doi:10.1371/journal.ppat.1000449

Cartwright EJ, Jackson KA, Johnson SD, Graves LM, Silk BJ, Mahon BE (2013) Listeriosis outbreaks and associated food vehicles, United States, 1998–2008. Emerg Infect Dis 19:1–9. doi:10.3201/eid1901.120393

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Cowan MM (1999) Plant products as antimicrobial agents. Clin Microbiol Rev 12:564–582

Cuaron JA, Dulal S, Song Y, Singh AK, Montelongo CE, Yu W, Nagarajan V, Jayaswal RK, Wilkinson BJ, Gustafson JE (2013) Tea tree oil-induced transcriptional alterations in *Staphylococcus aureus*. Phytother Res 27:390–396

de Mello JF, Einsfeldt K, Frazzon APG, da Costa M, Frazzon J (2008) Molecular analysis of the *iap* gene of *Listeria monocytogenes* isolated from cheeses in Rio Grande do Sul, Brazil. Braz J Microbiol 39:169–172

Delves-Broughton J (1990) Nisin and its application as food preservative. J Soc Dairy Technol 43(3):73–77

Duarte AL, de Menezes IRA, Braga MFBM, Leite NF, Barros LM, Waczuk EP, da Silva MAP, Boligon A, Rocha JBT, Souza DO, Kamdem JP, Coutinho HDM, Burger ME (2016) Antimicrobial activity and modulatory effect of essential oil from the leaf of *Rhaphiodon echinus* (Nees & Mart) Schauer on some antimicrobial drugs. Molecules 21:743. doi:10.3390/molecules21060743

Dussurget O, Dumas E, Archambaud C, Chafsey I, Chambon C, Hébraud M, Cossart P (2005) *Listeria monocytogenes* ferritin protects against multiple stresses and is required for virulence. FEMS Microbiol Lett 250:253–261

Edris AE (2007) Pharmaceutical and therapeutic potentials of essential oils and their individual volatile constituents: a review. Phytother Res 21:308–323

Fabri RL, Nogueira MS, Dutra LB, Bouzada MLM, Scio E (2011) Potencial antioxidante e antimicrobiano de espécies da família *Asteraceae*. Rev Bras Plant Med 13:183–189

Farber JM, Peterkin PI (1991) *Listeria monocytogenes*, a food-borne pathogen. Microbiol Rev 55:476–511

Fiorini F, Stefanini S, Valenti P, Chiancone E, de Biase D (2008) Transcription of the *Listeria monocytogenes fri* gene is growth-phase dependent and is repressed directly by Fur, the ferric uptake regulator. Gene 410:113–121

Fisher CW, Lee D, Dodge BA, Hamman KM, Robbins JB, Martin SE (2000) Influence of catalase and superoxide dismutase on ozone inactivation of *Listeria monocytogenes*. Appl Environ Microbiol 66:1405–1409

Girard D, Leclerqc A, Laurent E, Lecuit M, de Valk H, Goulet V (2014) Pregnancy-related listeriosis in France, 1984 to 2011, with a focus on 606 cases from 1999 to 2011. Euro Surveill 19:pii: 20909. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20909

Giuliano DA, Freire SE (2011) Nuevas secciones en *Baccharis* (*Asteraceae, Astereae*) de America del Sur. Ann Mo Bot Gard 98: 331–347

Gopal S, Berg D, Hagen N, Schriefer EM, Stoll R, Goebel W, Kreft J (2010) Maltose and maltodextrin utilization by *Listeria monocytogenes* depend on an inducible ABC transporter which is repressed by glucose. PLoS One 5(4):e10349. doi:10.1371/journal.pone.0010349

Gottleib OR, Taveira-Magalhães M (1960) Modified distillation trap. Chem Anal 49:114–115

Guedes RLM, Prosdocimi F, Fernandes GR, Moura LK, Ribeiro HAL, Ortega JM (2011) Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution. BMC Genomics. doi:10.1186/1471-2164-12-S4-S2

Haikarainen T, Papageorgiou AC (2010) Dps-like proteins: structural and functional insights into a versatile protein family. Cell Mol Life Sci 67:341–351

Hébraud M, Guzzo J (2000) The main cold shock protein of *Listeria monocytogenes* belongs to the family of ferritin-like proteins. FEMS Microbiol Lett 190:29–34

Imlay JA (2003) Pathways of oxidative damage. Annu Rev Microbiol 57: 395–418

Johnson DC, Dean DR, Smith AD, Johnson MK (2005) Structure, function, and formation of biological iron-sulfur clusters. Annu Rev Biochem 74:247–281

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. doi:10.1186/gb-2013-14-4-r36

Kim YG, Lee JH, Gwon G, Kim SI, Park JG, Lee J (2016) Essential oils and eugenols inhibit biofilm formation and the virulence of *Escherichia coli* O157:H7. Sci Rep 6:3637. doi:10.1038/srep36377

Körner H, Sofia HJ, Zumft WG (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. FEMS Microbiol Rev 27:559–592

Krawczyk-Balska A, Marchlewicz J, Dudek D, Wasiak K, Samluk A (2012) Identification of a ferritin-like protein of *Listeria monocytogenes*as a mediator of β-lactam tolerance and innate resistance to cephalosporins. BMC Microbiol 12:278. doi:10.1186/1471-2180-12-278

Kulisic T, Radonic A, Katalinic V, Milos M (2004) Use of different methods for testing antioxidative activity of orégano essential oil. Food Chem 85:633–640

Ledala N, Sengupta M, Muthaiyan A, Wilkinson BJ, Jayaswal RK (2010) Transcriptomic response of *Listeria monocytogenes* to iron limitation and Fur mutation. Appl Environ Microbiol 76:406–416

Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41: e108–e108

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) method. Methods 25:402–408

Lopez-Romero JC, González-Ríos H, Borges A, Simões M (2015) Antibacterial effects and mode of action of selected essential oils components against *Escherichia coli* and *Staphylococcus aureus*. Evid Based Complement Alternat Med 2015:795435. doi:10.1155/2015/795435

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550. doi:10.1186/s13059-014-0550-8

Lyon GJ, Novick RP (2004) Peptide signaling in *Staphylococcus aureus* and other Gram-positive bacteria. Peptides 25:1389–1403

Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJ (2001) Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. Microbiol Mol Biol Rev 65:44–79

Mazzarrino G, Paparella A, Chaves-López C, Faberi A, Sergi M, Sigismondi C, Compagnone D, Serio A (2015) *Salmonella enterica* and *Listeria monocytogenes* inactivation dynamics after treatment with selected essential oils. Food Control 50:794–803

McLaughlin HP, Xiao Q, Rea RB, Pi H, Casey PG, Darby T, Charbit A, Sleator RD, Joyce SA, Cowart RE, Hill C, Klebba PE, Gahan CG (2012) A putative P-type ATPase required for virulence and resistance to haem toxicity in *Listeria monocytogenes*. PLoS One 7(2): e30928. doi:10.1371/journal.pone.0030928

Mettert EL, Kiley PJ (2014) Coordinate regulation of the Suf and Isc Fe-S cluster biogenesis pathways by IscR is essential for viability of *Escherichia coli*. J Bacteriol 196:4315–4323

Milecka D, Samluk A, Wasiak K, Krawczyk-Balska (2015) An essential role of a ferritin-like protein in acid stress tolerance of *Listeria monocytogenes*. Arch Microbiol 197:347–351

Montero D, Bodero M, Riveros G, Lapierre L, Gaggero A, Vidal RM, Vidal M (2015) Molecular epidemiology and genetic diversity of *Listeria monocytogenes* isolates from a wide variety of ready-to-eat foods and their relationship to clinical strains from listeriosis outbreaks in Chile. Front Microbiol 6:384. doi:10.3389/fmicb.2015.00384

Miyamoto KN, Monteiro KM, da Silva CK, Lorenzatto KR, Ferreira HB, Brandelli A (2015) Comparative proteomic analysis of *Listeria monocytogenes* ATCC 7644 exposed to a sublethal concentration of nisin. J Proteome 119:230–237. doi:10.1016/j.jprot.2015.02.006

Negreiros MO, Pawlowski A, Zini CA, Soares GLG, Motta AS, Frazzon APG (2016) Antimicrobial and Antibiofilm activity of *Baccharis psiadioides* essential oil against antibiotic-resistant *Enterococcus faecalis* strains. Pharm Biol 54:3272–3279

Nes FD, Riboldi GP, Frazzon APG, d'Azevedo PA, Frazzon J (2010) Antimicrobial resistance and investigation of the molecular epidemiology of *Listeria monocytogenes* in dairy products. Rev Soc Bras Med Trop 43:382–385

Newton SM, Klebba PE, Raynaud C, Shao Y, Jiang X, Dubail I, Archer C, Frehel C, Charbit A (2005) The *svpA-srtB* locus of *Listeria monocytogenes*: *fur*-mediated iron regulation and effect on virulence. Mol Microbiol 55:927–940

Nguyen CC, Saier MH (1995) Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. FEBS Lett 377:98–102

Novick RP (2000) Pathogenicity factors and their regulation. In: Fischetti VA, Novick RP, Ferreti JJ, Portnoy DA, Rood JI (eds) Gram-positive pathogens. ASM, Washington, pp 392–407

O'Byrne CP, Karatzas KA (2008) The role of sigma B (sigma B) in the stress adaptations of *Listeria monocytogenes*: overlaps between stress adaptation and virulence. Adv Appl Microbiol 65:115–140

Olsen KN, Larsen MH, Gahan CG, Kallipolitis B, Wolf XA, Rea R, Hill C, Ingmer H (2005) The Dps-like protein Fri of *Listeria*

monocytogenes promotes stress tolerance and intracellular multiplication in macrophage-like cells. Microbiology 151:925–933

Pieta L, Garcia FB, Riboldi GP, de Oliveira LA, Frazzon APG, Frazzon J (2014) Transcriptional analysis of genes related to biofilm formation, stress-response, and virulence in *Listeria monocytogenes* strains grown at different temperatures. Ann Microbiol 64:1707–1714

Pieta L, Campos FS, Mariot RF, Prichula J, de Moura TM, Frazzon APG, Frazzon J (2015) Complete genome sequences of two *Listeria monocytogenes* serovars, 1/2a and 4b, isolated from dairy products in Brazil. Genome Announc 3(6): e01494-15. doi:10.1128/genomeA.01494-15

Py B, Barras F (2010) Building Fe-S proteins: bacterial strategies. Nat Rev Microbiol 8:436–446

Rea RB, Gahan CG, Hill C (2004) Disruption of putative regulatory loci in *Listeria monocytogenes* demonstrates a significant role for Fur and PerR in virulence. Infec Immun 72:717–727

Reichling J, Schnitzler P, Suschke U, Saller R (2009) Essential oils of aromatic plants with antibacterial, antifungal, antiviral, and cytotoxic properties—an overview. Forsch Komplementmed 16:79–90

Riedel CU, Monk IR, Casey PG, Waidmann MS, Gahan CG, Hill C (2009) AgrD-dependent quorum sensing affects biofilm formation, invasion, virulence and global gene expression profiles in *Listeria monocytogenes*. Mol Microbiol 71:1177–1189

Rodionov DA, Gelfand MS, Todd JD, Curson AR, Johnston AW (2006) Computational reconstruction of iron- and manganese-responsive transcriptional networks in alpha-proteobacteria. PLoS Comput Biol 2(12):e163. doi:10.1371/journal.pcbi.0020163

Santos JA, Alonso-García N, Macedo-Ribeiro S, Pereira PJB (2014) The unique regulation of iron-sulfur cluster biogenesis in a Gram-positive bacterium. Proc Natl Acad Sci USA 111:E2251–E2260

Simões CMO, Spitzer V (1999) Óleos voláteis. In: Simões CMO, Schenkel EP, Gosmann G, de Mello JCP, Mentz LA, Petrovick PR (eds) Farmacognosia: da planta ao medicamento, 6th edn. UFRGS, Porto Alegre, pp 387–416

Soni KA, Nannapaneni R, Tasara T (2011) The contribution of transcriptomic and proteomic analysis in elucidating stress adaptation responses of *Listeria monocytogenes*. Foodborne Pathog Dis 8:843–852

Swint-Kruse L, Matthews KS (2009) Allostery in the LacI/GalR family: variations on a theme. Curr Opin Microbiol 12:129–137

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc 7:562–578

Trémoulet F, Duché O, Namane A, Martinie B, Labadie JC, European Listeria Genome Consortium (2002) Comparison of protein patterns of *Listeria monocytogenes* grown in biofilm or in planktonic mode by proteomic analysis. FEMS Microbiol Lett 210:25–31

Upadhyay A, Upadhyaya I, Kollanoor-Johny A, Venkitanarayanan K (2013) Antibiofilm effect of plant derived antimicrobials on *Listeria monocytogenes*. Food Microbiol 36:79–89

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol Res:0034.1–0034.11

Vasconcelos JA, Deneer HG (1994) Expression of superoxide dismutase in *Listeria monocytogenes*. Appl Environ Microbiol 60:2360–2366

Williams T, Joseph B, Beier D, Goebel W, Kuhn M (2005) Response regulator DegU of *Listeria monocytogenes* regulates the expression of flagella-specific genes. FEMS Microbiol Lett 252:287–298

Xayarath B, Freitag NE (2012) Optimizing the balance between host and environmental survival skills: lessons learned from *Listeria monocytogenes*. Future Microbiol 7:839–852

Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: interactive pathway explorer. Nucleic Acids Res 39:W412–W415

Yang S, Xu H, Wang J, Liu C, Lu H, Liu M, Zhao Y, Tian B, Wang L, Hua Y (2016) Cyclic AMP receptor protein acts as a transcription regulator in response to stresses in *Deinococcus radiodurans*. PLoS One. doi:10.1371/journal.pone.0155010

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297

PLOS | NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

Genetic variability of *Taenia solium* cysticerci recovered from experimentally infected pigs and from naturally infected pigs using microsatellite markers

Mónica J. Pajuelo[1,2]*, María Eguiluz[1], Elisa Roncal[1], Stefany Quiñones-García[1], Steven J. Clipman[2], Juan Calcina[3], Cesar M. Gavidia[3], Patricia Sheen[1], Hector H. Garcia[1,4,5], Robert H. Gilman[2], Armando E. Gonzalez[3], Mirko Zimic[1], for the Cysticercosis Working Group in Peru[¶]

1 Laboratorio de Bioinformática y Biología Molecular, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru, **2** Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, United States of America, **3** School of Veterinary Medicine, Universidad Nacional Mayor de San Marcos, Lima, Peru, **4** Center for Global Health, Universidad, Peruana Cayetano Heredia, Lima, Peru, **5** Department of Microbiology, School of Science and Philosophy, Universidad Peruana Cayetano Heredia, Lima, Peru

¶ Membership of the Cysticercosis Working Group in Peru is provided in the Acknowledgments.
* monica.pajuelo.t@upch.pe

Citation: Pajuelo MJ, Eguiluz M, Roncal E, Quiñones-García S, Clipman SJ, Calcina J, et al. (2017) Genetic variability of *Taenia solium* cysticerci recovered from experimentally infected pigs and from naturally infected pigs using microsatellite markers. PLoS Negl Trop Dis 11(12): e0006087. https://doi.org/10.1371/journal.pntd.0006087

Editor: Andrea Winkler, Ludwig-Maximilians-University, UNITED STATES

Received: May 17, 2017

Accepted: October 31, 2017

Published: December 28, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was partially supported by the National Institutes of Health (grant numbers D43TW001140 and D43TW006581). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The adult *Taenia solium*, the pork tapeworm, usually lives as a single worm in the small intestine of humans, its only known definitive host. Mechanisms of genetic variation in *T. solium* are poorly understood. Using three microsatellite markers previously reported [1], this study explored the genetic variability of *T. solium* from cysts recovered from experimentally infected pigs. It then explored the genetic epidemiology and transmission in naturally infected pigs and adult tapeworms recovered from human carriers from an endemic rural community in Peru. In an initial study on experimental infection, two groups of three piglets were each infected with proglottids from one of two genetically different tapeworms for each of the microsatellites. After 7 weeks, pigs were slaughtered and necropsy performed. Thirty-six (92.3%) out of 39 cysts originated from one tapeworm, and 27 (100%) out of 27 cysts from the other had exactly the same genotype as the parental tapeworm. This suggests that the microsatellite markers may be a useful tool for studying the transmission of *T. solium*. In the second study, we analyzed the genetic variation of *T. solium* in cysts recovered from eight naturally infected pigs, and from adult tapeworms recovered from four human carriers; they showed genetic variability. Four pigs had cysts with only one genotype, and four pigs had cysts with two different genotypes, suggesting that multiple infections of genetically distinct parental tapeworms are possible. Six pigs harbored cysts with a genotype corresponding to one of the identified tapeworms from the human carriers. In the dendrogram, cysts appeared to cluster within the corresponding pigs as well as with the geographical origin, but this association was not statistically significant. We conclude that genotyping of microsatellite size polymorphisms is a potentially important tool to trace the spread of infection and pinpoint sources of infection as pigs spread cysts with a shared parental genotype.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Taenia solium, the pork tapeworm, is a major cause of epilepsy in developing countries. Although it has been deemed a potentially eradicable pathogen, it remains prevalent in rural communities. This two-part study aims to evaluate the utility of three microsatellite markers previously reported, to identify parasites and to establish relationships among them. In the first study, we evaluated the genetic variability of the progeny of two individual tapeworms by infecting two groups of three pigs each. We found variation of 8% and 0% in the two groups with respect to the parental tapeworm, indicating that the cysts source may be identifiable. Next, in the second study we described the genetic relationships among tapeworms obtained from four carriers and cysts obtained from eight naturally infected pigs in a rural community. We demonstrated that pigs can have two types of cysts, suggesting multiple infections. In addition, we found relatedness between 6 pigs and one tapeworm identified in the community. Our results indicate the potential for microsatellite markers to identify genetic relationships between parasites and thereby establish routes of transmission. It is likely that the limited number of microsatellites prevented us from establishing relatedness with more precision. Therefore, further evaluation of additional microsatellites is recommended.

Introduction

Taenia solium is a zoonotic parasite that affects both humans and pigs. Larval infection in the human brain results in neurocysticercosis, the sole main cause of acquired adult epilepsy in developing countries [2]. Neurocysticercosis is considered a potentially eradicable disease [3], and many efforts to implement interventions to control or eliminate this parasite are currently being explored [4, 5]; however, there are some aspects of the epidemiology, such as distribution and transmission of the parasite among endemic communities that are still unknown. Control strategies for cysticercosis vary in scale and scope. Understanding the interaction between the parasite, swine host, and humans by observing transmission dynamics and clustering of infection through the use of genotyping can help determine the most cost-effective intervention for the given setting.

Although the genetic variation of *T. solium* has not yet been fully described, a detailed understanding of *T. solium* population genetic structure is vital to determining the transmission and other epidemiological features of this disease [6–8]. Moreover, the study of genetic variants may help elucidate the reproduction aspects of the parasite [8].

Genotyping tools are useful to study the etiology and distribution of parasites among populations [9, 10]. The genetic variability of *T. solium*, in particular, demonstrates geographic distribution that has been previously described in the literature [11–13]. In experimental infection, genetic variability has been shown among cysts originating from a given parental tapeworm, as well as an association between the genotypes of the parental tapeworm and its cysts progeny [14].

Given that *T. solium* is a monoecious self-fertilizing parasite, most common genetic markers show poor genetic variability, and therefore genetic markers with higher mutations rates, such as microsatellites are needed. Microsatellites have been used in other cestodes, like *Echinococcus multilocularis*, in which the, spatial distribution of strains was studied in different regions of the world [9]. In particular, this was analyzed at a local scale, in the French Ardennes area, Italy, Hungary, and Norway [15–18]. Remarkably, it was found that the use of

the single microsatellite marker, EmsB, is sufficient for molecular tracking of transmission [19, 20]. Albeit not from public health perspective, the use of microsatellites has also been reported in the study of genetic and reproductive aspects of the cestode *Schistocephalus solidus* [21], as well as, in the Mendelian inheritance in *Oochoristica javaensis*[22].

Recently, our group has sequenced the whole genome of two *T*. *solium* isolates. After identifying more than 9,000 microsatellite sequences along the complete genome, we evaluated 36 microsatellites, and identified a set of seven polymorphic microsatellite markers with potential use for genotyping [1]. Using those markers, strains from communities in the North and South of Peru were able to be discriminated, and a small amount of intra-community genotypic variability was able to be detected [1].

This two-pronged investigation including one experimental and one epidemiologic study was developed to evaluate *T*. *solium* genetic variability in a public health context. The experimental portion of this study aimed to establish empirical evidence that offspring cysts have the same genotype as the parental tapeworm. We evaluated the genotypes of offspring cysts derived from pigs experimentally infected with *T*. *solium* proglottids of the same tapeworm. Experimental evidence was designed to inform the second part of this study, and is important in the ability to pinpoint sources of infection and allow for epidemiologic tracing. The epidemiologic study aimed to examine associations between tapeworms and cysts found in pigs. We analyzed the association of genotypes of tapeworms from human carriers and cysts from naturally infected pigs in a rural community in Piura, a northern city of Peru to provide an initial exploratory assessment of the epidemiology of transmission in this community.

Materials and methods

Genetic variation in cysts developing in pigs infected with eggs from the same tapeworm

The objective of this experiment was to identify and evaluate the spontaneous variability of cyst genotypes compared to the genotypes of the parental tapeworm from which they originated. For this, two groups of three pigs each were experimentally infected with proglottids of two different tapeworms with distinct known genotypes based on four microsatellite markers.

Tapeworm samples. The *T*. *solium* samples were donated from the repository of the Cysticercosis Working Group in Peru, and they were obtained from "complete" tapeworms (e.g. portion very close to the scolex and gravid proglottids) expelled as residual samples after routine treatment of two patients, one from Apurímac in the South Highlands of Peru (Tapeworm TA) and one from Cajamarca in the North Highlands of Peru (Tapeworm TB). Gravid proglottids were stored in 25% glycerol supplemented with penicillin (1000 UI/mL), gentamicin (100 μg/mL), amphotericin B (0.02 mg/mL), and streptomycin (1 mg/mL) at 4˚C until infection. The proximal portion next to the scolex was stored in absolute ethanol at room temperature until genotyping was performed. Both had different genotypes, evaluated by sequencing (described below in the Genotyping section). The closest segment to the scolex portion of the parasite was used for DNA extraction and genotyping the tapeworm, since it does not have proglottids with eggs [23]. The gravid proglottids, which proved to be viable (see below), were used to experimentally infect pigs.

Animals. Six one-month-old female piglets (*Sus scrofa domestica*) were obtained from a farm free of *T*. *solium* in Lima, Peru. Animals were confirmed to be negative to the antibody detection assay by an enzyme-linked immunoelectrotransfer blot (EITB, western blot) assay using lentil-lectin purified parasite glycoprotein antigens [24].

Viability of oncospheres. To proceed to the artificial infection of pigs, it was important to know if the oncospheres contained in the proglottids were viable. This was done as described

by Verástegui et al [25]. Briefly, the oncospheres were released using hypochlorite at 0.75% for 10 minutes to weaken the egg layer. Subsequently they were washed 3 times with RPMI 1640 medium, 0.4% trypan blue added. It was observed under the microscope and the oncospheres that excludes the dye were considered viable.

Infection of pigs. Three piglets were infected with one proglottid from tapeworm TA each, and the other three with proglottids from tapeworm TB. Infection was carried out seven months after tapeworms were obtained.

Pigs were identified with a numbered ear-tag. Proglottids were administered inside a piece of banana as previously described [26]. Infection was confirmed by a positive EITB Western blot test [24] two weeks after infection.

Porcine blood sampling. Blood was obtained from the animals for cysticercosis serology two weeks after infection. A trained veterinarian collected the blood sample from the anterior cava vein using vacuum tubes; serum was isolated by centrifugation at 3500 r.p.m. for 5 minutes. Samples were stored at -20˚C.

Necropsy. All pigs were euthanized 7 weeks post infection. Each pig was injected with Ketamine (20 mg/kg) and Xylacine (2 mg/kg) IM to produce sedation. Under sedation, the pigs were injected with sodium pentobarbital IV to produce euthanasia. Necropsy was performed immediately after euthanasia [26]. Full carcass dissections were performed. Healthy cysts were recovered from each carcass.

Collection of cysts for DNA extraction. Healthy cysts were individually and randomly selected and collected from the trunk, muscles from legs and arms, brain and heart, if available for each pig. Cysts were washed with saline solution 0.9% twice and a final wash with ethanol, so the sample is free from porcine tissue. Then each cyst was stored in absolute ethanol at 4˚C. **A healthy cyst was defined** as a sack containing a transparent clear fluid and a white structure called the scolex. Cysts in this form are known to be mostly viable [27].

Cyst DNA purification. DNA purification was done using the QIAmp DNA Mini Kit (QIAGEN, Hilden, Germany) according to manufacturer´s instructions. The DNA quality and quantity was verified by UV spectrometry (Nanodrop 2000c) [28] and was stored at -20˚C until use.

Microsatellite genotyping. Four microsatellite markers were used in this study: TS_S SR09, TS_SSR27, TS_SSR28, and TS_SSR32. Other markers reported in our previous study were excluded due to the following reasons: TS_SSR01 had shown low polymorphism among northern strains [1]. TS_SSR16 and TS_SSR18 were dinucleotides, with a high risk of unprecise assessment of polymorphism. Sanger sequencing was performed to assess size polymorphisms of the microsatellites. This technique causes loss of information at the 5´ end, which could include the repetitive motif of the microsatellite, therefore we designed a new set of external primers for the specific microsatellites SSR09, SSR27, SSR28, and SSR32 respectively: TS_SSR09-F (5´- TGGCATTCGACTGGATGACC -3´), TS_SSR09-R (5´- AGAGAAG CAACAGAATACTGC -3´), TS_SSR27-F (5´- AGGTAGACCACCTCCGTCTC -3´), TS_SSR 27-R (5´- GGAAATTCGCATGGCTGTGG -3´), TS_SSR28-F (5´- TCTACCCCGTCAGTTGAG GT -3´), TS_SSR28-R (5´- GGTGTGAATTAACCAGCTAG -3´), TS_SSR32-F (5´- GGATGT GACGGGGTTTGACA -3´), and TS_SSR32-R (5´- CATTAGGGGTTCAGTCGGGG -3´) using Primer3 [29, 30]. The PCR reaction volume (25 μL) consisted of: Buffer 1X (Invitrogen), 2 mM $MgCl_2$, 0.2 mM dNTPs each one, Forward primer: 1μM, Reverse primer: 1 μM and 20 ng of DNA, Taq polymerase 0.3 U. PCR was conducted in a MJ Research MiniCycler PTC-150 thermocycler with a hot cover using the following temperature profile except where otherwise stated: the initial denaturation step was at 95˚C for 5 minutes, followed by 35 cycles of 95˚C for 45 seconds, 62˚C for 45 seconds and 72˚C for 2 minutes and a final extension at 72˚C for 5

minutes. PCR products were sent for DNA sequencing at Macrogen Corp., Rockville MD. The size of the microsatellite markers was calculated using a reference sequence obtained in our previous study, for comparison [1]. We verified that the variation in size was due to the number of repeats and not due to any other mutation in the flanking regions of the microsatellite marker.

Analysis of cysts found in naturally infected pigs

The aim of this experiment was to evaluate the genetic variability of *T*. *solium* cysts from pigs recovered in a natural environment and its association with tapeworm carriers.

Design. First, a mapping and census of all houses was done in the community including GPS locations of each house. Tapeworm carriers had been previously identified three months prior in a previous study [31] and tapeworms were obtained from that study. Infected pigs were detected by tongue examination [32]. These animals were euthanized, and cysts were randomly recovered from the entire carcasses. Tapeworms and cysts were genotyped using the four microsatellites markers described above [1].

Study site. Pampa Elera Baja is a rural community located in the highlands of the Northern Region of Piura, in Peru. It has about 700 inhabitants. The prevalence of *T*. *solium* taeniosis was estimated in 1% [31]. 79/170 (42%) of families raised pigs at small scale.

Mapping and census. Mapping and a census of all houses in the community was performed. Mapping included geographic location (latitude and longitude coordinates) of each house recorded using global positioning system (GPS) receivers (GeoExplorer CE XT; Trimble, Westminster, CO). For the census, we obtained additional information from each person residing in the house, including age, sex, origin and characteristics of the house, such as material of construction, presence of latrine, disposal of feces, source of water, and treatment of water for consumption. GPS locations of houses that own pigs were used for analysis of association between genetic distances and geographic distances.

Tongue examination. Tongue examination was performed to identify possibly infected pigs [32]. Tongue examination is 100% specific [32]. The tongue of each pig was held using a special forceps while a trained veterinarian conducted a manual examination starting at the base of the tongue and palpating down to the tip to detect the presence of nodules and cysts. Pigs were considered positive if cysts were observed or palpated in the tongue muscle or base, otherwise the pig was considered negative.

Necropsy. Pigs were purchased from their owners and moved to a separate euthanasia area. Necropsy was performed as explained in the previous section. Healthy cysts were recovered from each carcass for microsatellite genotyping.

Microsatellite genotyping. DNA from tapeworms recovered from human carriers was previously obtained by our group (Watts, 2014) 3 months before, therefore no important variations at the population level are expected. DNA from cysts was collected from naturally infected pigs. All DNA samples collected were used for genotyping. All tapeworms and 10 to 14 cysts per pig were processed for DNA extraction, DNA purification, PCR amplification, and sequencing for microsatellites TS_SSR09, TS_SSR27, and TS_SSR28 as explained in previous sections.

Statistical genetic analysis. Expected heterozygosity was estimated with Arlequin V.3.5 software [33]. To determine the genetic distances between tapeworms and the cysts from different pigs, pairwise Nei's genetic distance(Da) was calculated using POPULATIONS software version 1.2.32 [34]. A UPGMA (unweighted pair group method with arithmetic mean) dendrogram was inferred and trees were constructed using the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

The geographic distances were calculated using data from the GPS records and the on-line distance calculator: Movable Type Scripts (http://www.movable-type.co.uk/scripts/latlong.html). The geographic location of the pigs was considered identical to the geographical location of the owner household. The associations between the genetic distances (Da) and geographic distances were tested by the Mantel test [35] using R [36].

The probability that a naturally infected pig had different types of cysts in proportions due to the natural variation was calculated by the binomial test.

Ethics statement

All procedures complied and were **approved** by the Ethics Committee for Animal Use (CIEA: Assurance NumberA5146-01) at Universidad Peruana Cayetano Heredia (Lima, Peru) under Protocol Number 62400 for the experimental infection, and Protocol Number 61340 for the field study. *T. solium* tapeworms used for the experimental infections were donated from the repository of the Cysticercosis Working Group in Peru, and had been obtained from tapeworms expelled as residual samples after routine treatment. *T. solium* tapeworms used in the field study were obtained from a previous study conducted in the community of Pampa Elera [31].

Results

Genetic variation in cysts developing in pigs infected with eggs from the same tapeworm

Before infection, viability was established, Tapeworm TA had a viability of 68% and Tapeworm TB had a viability of 62%. All the pigs were successfully infected, as confirmed by necropsy. We processed a total of 39 healthy cysts from Tapeworm TA and 27 healthy cysts from Tapeworm TB. The two tapeworms showed different genotypes for all markers, except for SSR32 (Table 1).

Microsatellite SSR32 was found to be monomorphic among tapeworms TA and TB, and among cysts from different pigs (results are shown in S1 Table); therefore, it was not included in the analysis. Based on sequencing results, we defined that two genotypes were the same if the size bands in the three loci for SSR09, SSR27, and SSR28 were identical. Thirty six out of 39 examined cysts (92.3%, 95%CI: [79.1% - 98.4%]) from Tapeworm TA and 27 out of 27 examined cysts (100%, 95% CI: [87.2%-100%]) from Tapeworm TB showed the same genotype as the parental tapeworm (Table 2, and S1 Table). There are spontaneous mutations that appear to occur naturally and are not higher than two repeats (6 nucleotides). Therefore, the probability that a cyst with a difference of more than two repeats be a spontaneous evolution of the parental tapeworm is 7.7%. In analyzing each individual marker, 1/39 cysts varied in any of the markers (2.6%, 95% CI [0.1%-13.5%]). The results of the markers that differed from the parental tapeworm were amplified and sequenced twice and the differences were consistently observed.

Table 1. Genotypes of tapeworms used to experimentally infect pigs obtained by sequencing.

	SSR09 (GGT)	SSR27 (GAA)	SSR28 (GTA)	SSR32 (AGC)
Tapeworm TA	169	168	224	176
Tapeworm TB	160	153	221	176

https://doi.org/10.1371/journal.pntd.0006087.t001

Table 2. Genotype of cysts from experimental infection based on sequencing.

Pig	Number of cysts found at necropsy	Number of examined cysts	SSR09	SSR27	SSR28
A1	1245	10	169	168	224
		1	169	168	**218**[a]
		1	169	**165**[a]	224
A3	499	13	169	168	224
		1	**163**[a]	168	224
A7	1611	13	169	168	224
B4	675	11	160	153	221
B5	11	5	160	153	221
B6	135	11	160	153	221

Pigs A1, A3 and A7 were infected with Tapeworm TA proglottids and Pigs B4, B5 and B6 were infected with Tapeworm TB proglottids.

[a] Alleles that were different (in size) from the original tapeworm allele

https://doi.org/10.1371/journal.pntd.0006087.t002

Analysis of cysts found in naturally infected pigs

Census and mapping. Pampa Elera Baja is a typical rural community with absence of sanitary facilities, where pigs roam freely to forage for food. A total of 530 people were surveyed in 170 houses in the community. Two hundred and seventy one (51.2%) were women and 259 (48.8%) were men. Median age of the population was 26 years old (IQR 11–45). This is a rural community where 90% of the houses were made of regional material (wood and mat). Ninety four percent of the population consumed water from river or ditch, 74% reported treatment of water for consumption, and 20% consumed untreated water. Eighty eight percent of the households did not have a latrine, and 6% had an artisanal latrine. Eighty six percent of the population reported to defecated in the field, while only 8% did in a latrine. Regarding porcine husbandry, a total of 79 families (46%) declared that they raised pigs. Families raised 3.5 pigs on average, ranging from 1 to 12 per family. The map of the village, showing the location of infected pigs and tapeworm carriers identified in a previous study [31] is shown in Fig 1.

Among the four tapeworm carriers, two of them (3 and 4) did not have a latrine and declared that they defecated in the field. The other two tapeworm carriers (1 and 2) declared that they had a latrine and used them.

Identification of infected pigs. During field work, a total of 303 pigs were evaluated, more than what was declared during census. A total of 9 tongue positive pigs were identified, of which 8 were slaughtered and healthy cysts were collected per animal. All infected pigs were born and raised in Pampa Elera Baja, they all roamed freely as stated by the owner. The characteristics of the animals are described in Table 3.

Genotyping. Four microsatellites markers were used to genotype tapeworms and cysts: SSR09, SSR27, SSR28, and SSR32. Results are shown in Tables 4 and 5, and S1. Because SSR32 was found to be monomorphic, we excluded it from the analysis (S1 Table). Based on sequencing results, we defined that two genotypes were the same if the size bands in the three loci for SSR09, SSR27, and SSR28 were identical. Accordingly, six pigs (P1-P6) had cysts with a genotype identical to tapeworm T3. Two pigs (P7 and P8) had cysts with genotypes not found in any of the tapeworms; both of them were from the same household (#40) (Table 3).

Two pigs (P4 and P5), that have cysts with genotype matching the genotype of tapeworm T3, lived 220 m away from that tapeworm. The pig P6 that had cysts that matched with

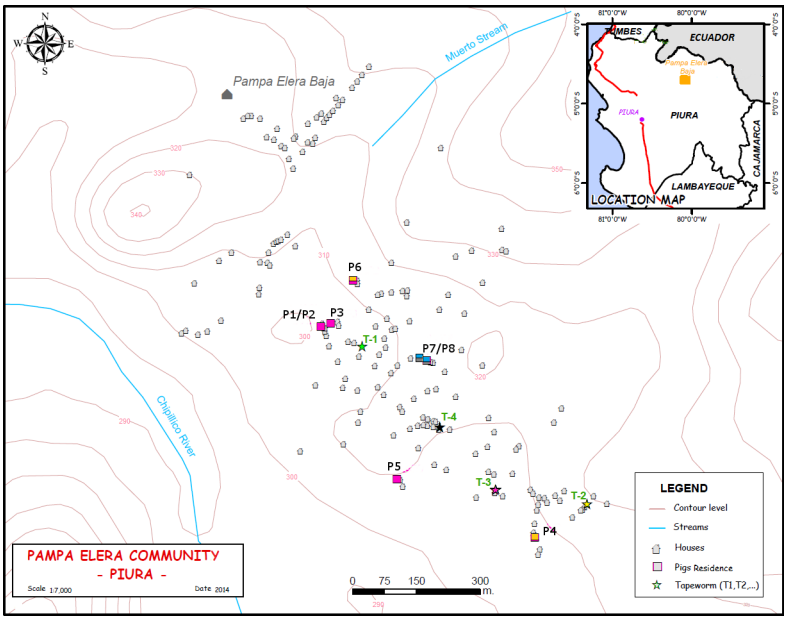


Fig 1. Map of the community showing the location of tapeworm carriers and cysticercosis-positive pigs. This figure was created using ArcGIS and http://escale.minedu.gob.pe/descargas/mapa.aspx was used also for base layer. Colors represent the genotype of tapeworms and cysts in pigs.

https://doi.org/10.1371/journal.pntd.0006087.g001

tapeworm T3 is the pig that lives furthest away (590 m) from tapeworm T3, other pigs live within that area.

In the field study, cysts recovered from the same pig showed some genetic variation. Four pigs harbored cysts with a unique genotype; however four pigs harbored cysts with two different genotypes, in proportions that ranged from 10% to 43% (Table 5). Based on the spontaneous variability found in the experimental infection, the likelihood that a pig had two different types of cysts in proportions of about 40%/60% is not likely to occur as a natural spontaneous variation ($P<0.0001$). Therefore we believe that at least two pigs: P8 and P6 harbored cysts with different parental origin (i.e. multiple infections). Hence, 32% of cysts recovered from pig P6 and 35% of cysts recovered from pig P8 are likely to be of different parental origin. The expected heterozygosity ranged from 0 to 0.34 (S2 Table).

Dendrogram analysis

The UPGMA tree (based on distances computed as differences of the number of "repeated short sequences"), of 8 groups of cysts recovered from different pigs, and four tapeworms based on genetic distances were classified into two groups associated with two regions of different altitude, suggesting that the genetic diversity was related to the geographical location (Fig 2). One group is comprised of the cysts obtained from pigs (P1, P2, P3, and P5) and tapeworm T1 and T3. The second group includes cysts from pigs (P4, P6, and P8), and tapeworms T2 and T4.

To analyze the association of genetic variability and geographical distance, the Mantel test was used for the microsatellite markers. No significant correlation between genetic and geographic distances was found (Spearman Rank correlation coefficient $r = 0.166$, Mantel $P = 0.46$).

Discussion

In this study we show important evidence of genetic variability of *T. solium*, and present a promising genotyping tool based on DNA microsatellite markers. First, our data shows that

Table 3. Characteristics of pigs with cysticercosis used in this study.

Pig	Sex	Age (months)	House	Number of cysts found (counted)	Number of cysts evaluated
P1	Male	7	117	2276	15
P2	Male	7	118	959	13
P3	Male	7	13	1837	12
P4	Female	9	79	634	10
P5	Male	9	91	1448	10
P6	Female	12	1	65	10
P7	Female	24	40	1064	12
P8	Female	24	40	114	14

https://doi.org/10.1371/journal.pntd.0006087.t003

the probability that a cyst has the same genotype (i.e. same microsatellite marker sequence) as the parental tapeworm in an experimental infection is 92.3%. Second, naturally infected pigs in an endemic rural community harboring subtypes of cysts with different genotypes, could be explained by multiple infections. This tool is likely to correctly identify the progeny of a tapeworm.

Defining a genotype as a unique combination of the microsatellite markers, the experimental infection study showed that the progeny of a given tapeworm had an almost identical genotype as the parental worm. This was specially noted in the cysts originating from the tapeworm TB from Cajamarca, in the north of Peru, where all sequenced cysts had exactly the same genotype. This suggests a low genetic variability per generation that can be explained by *T*. *solium* being a monoecious parasite that self-fertilizes [23, 37]. Through our methodology, we can be assured that the cysts (3/39) that displayed different genotypes from the parental tapeworm are due to spontaneous mutations of the microsatellites (e.g. slippage) [38, 39] and not due to experimental PCR, since markers that resulted in different band size by sequencing were genotyped twice, including the PCR reaction itself. The spontaneous mutations in cysts originating from the same tapeworm occurred at a rate of 7.7%, varying less than 2 repeats per generation. Not considering this natural spontaneous genetic variability of cysts, could cause overestimation of multiple infections in pigs; however, higher proportions of different genotypes may point to multiple infections. Despite this natural variation, we show lower genetic variability of *T*. *solium* than previously reported, using RAPD test [14] in an experimental infection. This discrepancy could be explained by the fact that the RAPD test often yields higher variability due to its random nature opposed to a DNA sequencing approach.

The epidemiologic study confirmed genetic variability among cysts obtained from pigs naturally infected within the community, as previously reported [11, 12]. One aim of this study was to show evidence of transmission between tapeworms and cysts using microsatellite

Table 4. Genotype of tapeworms found in Pampa Elera community.

Tapeworm	House	Genotype		
		SSR09	SSR27	SSR28
T1	16	160	153	215
T2	69	160	156	218
T3	83	157	153	215
T4	94	160	153	218

Band size determined by sequencing.

https://doi.org/10.1371/journal.pntd.0006087.t004

Table 5. Genotypes of cysts excised from naturally infected pigs based on sequencing.

Pig		Genotype			
	Number of examined cysts	SSR09	SSR27	SSR28	Genotype proportion
P1	13	157	153	215	1.00
	1	[a]	153	215	[b]
	1	[a]	[a]	215	[b]
P2	12	157	153	215	1.00
	1	[a]	153	215	[b]
P3	11	157	153	215	1.00
	1	157	153	[a]	[b]
P4	9	160	153	221	0.90
	1	157	153	215	0.10
P5	8	157	153	215	1.00
	2	[a]	[a]	215	[b]
P6	6	160	153	221	0.60
	4	157	153	215	0.40
P7	9	160	156	215	0.82
	2	160	156	221	0.18
	1	160	[a]	215	[b]
P8	8	160	156	215	0.57
	6	160	156	221	0.43

[a] Not enough DNA.

[b] Not considered for genotype proportion calculation.

https://doi.org/10.1371/journal.pntd.0006087.t005

markers; important information was obtained from the analysis of the genotypes of cysts obtained from 8 pigs and 4 tapeworms in the community, as explained below.

Two pigs (P7 and P8) from the community had cysts with two microsatellite genotypes different from the identified tapeworms. Since all sacrificed pigs were reported to be born and raised in Pampa Elera, it is likely that they became infected from unidentified tapeworm carriers. Also 23% of people surveyed came from different communities in Piura, Tumbes, and some from Lima. Therefore, it is possible that infections from transient tapeworm carriers may have occurred as well.

Six pigs harbored cysts that matched the genotype of tapeworm T3 (P1-P6), showing some degree of concordance between the tapeworm and cysts recovered from pigs (also shown in the dendrogram); however, the Mantel test did not show a significant association between genetic distances and geography. Nevertheless, it has been shown that pigs can roam several kilometers away from their homes [40, 41]. At this point, it is possible to state that: two or more tapeworms in the community could have the same genotype, and/or the low number of microsatellite markers used does not allow for capture of a finer overall genotype. Genotypes of Tapeworms T1 and T2 did not appear among the excised cysts, the corresponding tapeworm carriers declared having and using a latrine for defecation.

Four pigs harbored one type of cyst and four other pigs harbored two types of cysts. The fact that cysts of one genotype were found within an individual pig may possibly be explained by acquired or protective immunity. The finding of cysts with two genotypes in the proportions found in this study suggest either multiple infections (animals that have eaten food contaminated with two different tapeworms either at the same time or at different times, i.e. reinfection), that the tapeworm had an intrinsic source of variability such as recombination

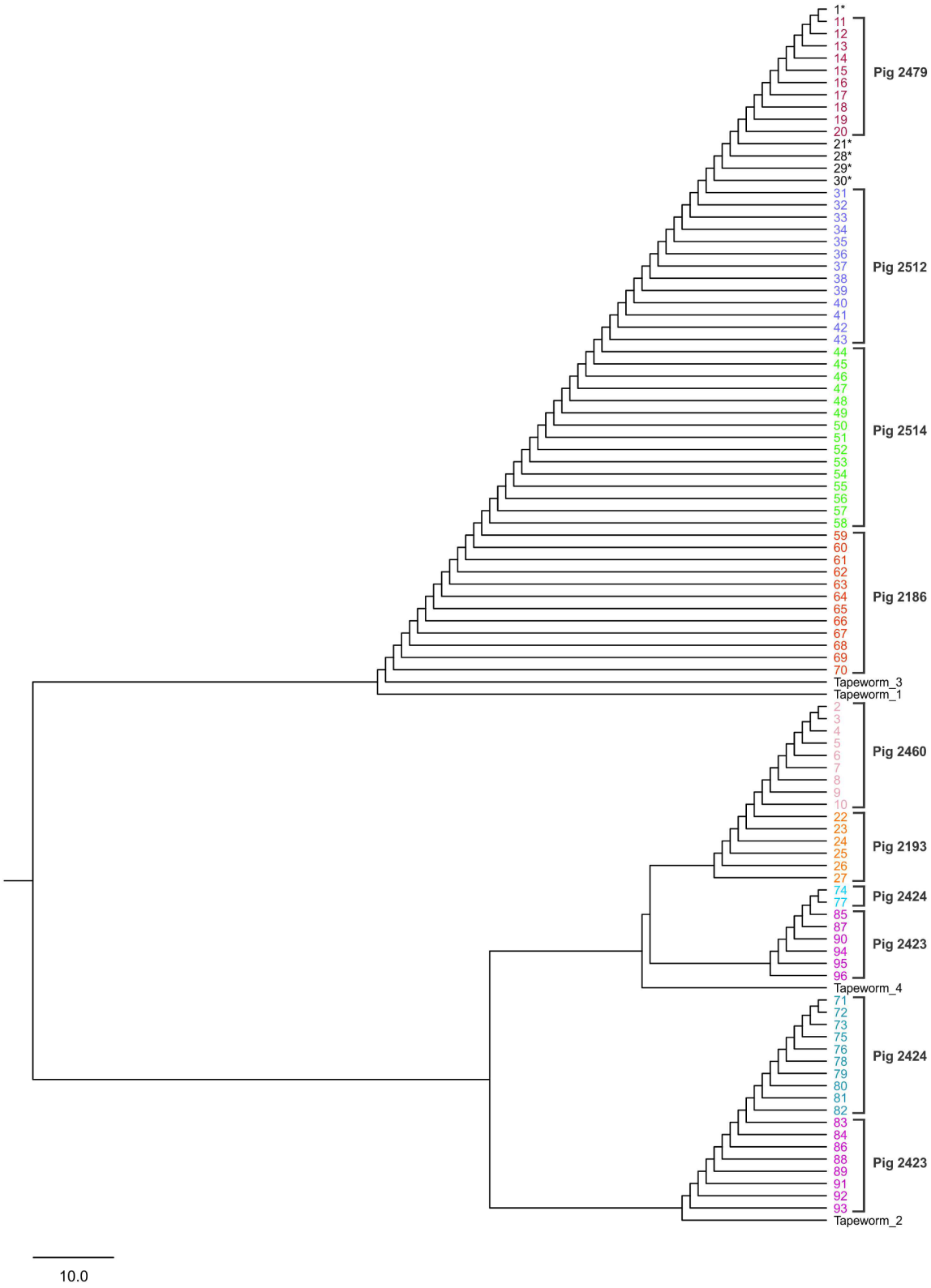


Fig 2. UPGMA dendrogram depicting Dc genetic distances between 96 cysts based on three polymorphic nuclear microsatellite loci. 1–10 (Pig P4), 11–20 (Pig P5), 21–30 (Pig P6), 31–43 (Pig P2), 44–58 (Pig P1), 59–70 (Pig P3), 71–82 (Pig P7), 83–96 (Pig P8). Two main groups are shown.

https://doi.org/10.1371/journal.pntd.0006087.g002

(for instance due to cross fertilization), or that a tapeworm carrier was infected with more than one tapeworm.

Pigs with only one type of cyst were all male and 7 to 9 months old. On the other hand, all pigs that harbored two types of cysts were female, in whom there could be a drop in immunity peri-partum [42], which could favor reinfection. Also, De Aluja *et.al*. (1999) reported that experimentally infected pigs were resistant to reinfection for at least five months [43], after which they could become re-infected. All pigs that harbored two types of cysts were older than 9 months old, possibly allowing enough time to be infected twice. Also, it has been reported that pigs from 7 to 9 months old roam greater distances from the household [41], therefore the chances of becoming infected from other sources further from the pigs´ households could be higher. Finally, it was found that two pigs (P7 and P8) that live in the same household had the same two types of cysts. It has been reported by Pray *et. al*. (2016) that pigs from the same household have identical roaming ranges and stay together as a herd, possibly explaining why the genotypes were found to be the same.

It is likely that we missed infected pigs and tapeworm carriers. In our previous study [31], in which tapeworm carriers were identified, we invited people from the community and not all were willing to participate. Also, we may have missed tapeworm carriers due to use of an imperfect diagnostic tool (spontaneous sedimentation in tube technique and microscopy) [31]. Regarding pigs, only tongue positive pigs were analyzed, thus it is possible that more infected pigs harboring different cysts are present in the community.

In this study, we initially tested four previously reported polymorphic microsatellite markers; however, based on sequencing analysis, one of them, TS_SSR32, was found to be monomorphic. This finding reinforces conceptions about the lack of precision in the automated capillary system to identify size polymorphisms, as also reported by Manrique et al. [44]. Therefore, we evaluated the genotype of cysts using three microsatellite markers. This number may seem small; however, another study found that a single multilocus microsatellite (the EmtB) of *Echinococcus multilocularis*, was informative enough to genotype strains with a high enough resolution to track transmission [19]. Therefore, it is likely that the three selected microsatellites are sufficient. However further studies are required to evaluate this in a larger population and to include additional microsatellite markers to improve precision to establish relatedness among individuals within a community.

In conclusion, this study shows that these microsatellite markers constitute a promising tool, and further genotyping should be done using additional markers to explore their potential in tracing transmission in an endemic community.

Supporting information

S1 Table. Genotypes of evaluated cysts based on sequencing.
(DOCX)

S2 Table. Expected heterozygosity of microsatellite markers by group of cysts genotyped from each pig in the rural community.
(DOCX)

Acknowledgments

We want to acknowledge Lauralee Woods and Katherine Murray, Peace Corps members for their enormous support in contacting the population of Pampa Elera in Piura and to the Pampa Elera community for their collaboration. We also want to thank Eng. Carmen Gamero Huamán for creating the map presented in this article, and to the members of the Cysticercosis

Working Group in Peru, especially to Dr. Manuela Verástegui, Daniela Alvarez and the Center for Global Health for their collaboration in Piura.

Author Contributions

Conceptualization: Mónica J. Pajuelo, Hector H. Garcia, Robert H. Gilman, Mirko Zimic.

Data curation: María Eguiluz, Elisa Roncal.

Formal analysis: Mónica J. Pajuelo, María Eguiluz, Steven J. Clipman.

Funding acquisition: Hector H. Garcia, Robert H. Gilman.

Investigation: Mónica J. Pajuelo, María Eguiluz, Elisa Roncal, Stefany Quiñones-García, Steven J. Clipman, Juan Calcina.

Methodology: Mónica J. Pajuelo, Patricia Sheen, Hector H. Garcia, Robert H. Gilman, Armando E. Gonzalez.

Resources: Robert H. Gilman, Mirko Zimic.

Supervision: Mónica J. Pajuelo, Cesar M. Gavidia, Armando E. Gonzalez, Mirko Zimic.

Validation: Elisa Roncal, Patricia Sheen.

Writing – original draft: Mónica J. Pajuelo, María Eguiluz, Hector H. Garcia.

Writing – review & editing: Mónica J. Pajuelo, María Eguiluz, Elisa Roncal, Stefany Quiñones-García, Steven J. Clipman, Juan Calcina, Cesar M. Gavidia, Patricia Sheen, Hector H. Garcia, Robert H. Gilman, Armando E. Gonzalez, Mirko Zimic.

References

1. Pajuelo MJ, Eguiluz M, Dahlstrom E, Requena D, Guzman F, Ramirez M, et al. Identification and Characterization of Microsatellite Markers Derived from the Whole Genome Analysis of *Taenia solium*. PLoS Negl Trop Dis. 2015; 9(12):e0004316. Epub 2015/12/25. https://doi.org/10.1371/journal.pntd.0004316 PMID: 26697878; PubMed Central PMCID: PMCPmc4689449.
2. Singh G, Burneo JG, Sander JW. From seizures to epilepsy and its substrates: neurocysticercosis. Epilepsia. 2013; 54(5):783–92. https://doi.org/10.1111/epi.12159 PMID: 23621876
3. Center for Disease Control and Prevention. Recommendations of the International Task Force for Disease Erradication. Atlanta, Georgia, USA: Public Health Service, U.S. Department of Health and Human Services, CDC; 1993 RR-16.
4. Garcia HH, Gonzalez AE, Rodriguez S, Gonzalvez G, Llanos-Zavalaga F, Tsang VC, et al. Epidemiology and control of cysticercosis in Peru. Revista peruana de medicina experimental y salud publica. 2010; 27(4):592–7. PMID: 21308201
5. Garcia HH, Gonzalez AE, Tsang VC, O'Neal SE, Llanos-Zavalaga F, Gonzalvez G, et al. Elimination of *Taenia solium* Transmission in Northern Peru. N Engl J Med. 2016; 374(24):2335–44. Epub 2016/06/16. https://doi.org/10.1056/NEJMoa1515520 PMID: 27305193.
6. Nakao M, Okamoto M, Sako Y, Yamasaki H, Nakaya K, Ito A. A phylogenetic hypothesis for the distribution of two genotypes of the pig tapeworm *Taenia solium* worldwide. Parasitology. 2002; 124(Pt 6):657–62. PMID: 12118722
7. Shrivastava J, Qian BZ, McVean G, Webster JP. An insight into the genetic variation of *Schistosoma japonicum* in mainland China using DNA microsatellite markers. Molecular ecology. 2005; 14(3):839–49. https://doi.org/10.1111/j.1365-294X.2005.02443.x PMID: 15723675
8. Campbell G, Garcia HH, Nakao M, Ito A, Craig PS. Genetic variation in *Taenia solium*. Parasitology international. 2006; 55 Suppl:S121–6. https://doi.org/10.1016/j.parint.2005.11.019 PMID: 16352464
9. Knapp J, Bart JM, Glowatzki ML, Ito A, Gerard S, Maillard S, et al. Assessment of use of microsatellite polymorphism analysis for improving spatial distribution tracking of *Echinococcus multilocularis*. J Clin Microbiol. 2007; 45(9):2943–50. Epub 2007/07/20. https://doi.org/10.1128/JCM.02107-06 PMID: 17634311; PubMed Central PMCID: PMCPMC2045259.

10. Criscione CD, Valentim CL, Hirai H, LoVerde PT, Anderson TJ. Genomic linkage map of the human blood fluke *Schistosoma mansoni*. Genome biology. 2009; 10(6):R71-2009-10-6-r71. Epub Jun 30. https://doi.org/10.1186/gb-2009-10-6-r71 PMID: 19566921

11. Maravilla P, Souza V, Valera A, Romero-Valdovinos M, Lopez-Vidal Y, Dominguez-Alpizar JL, et al. Detection of genetic variation in *Taenia solium*. The Journal of parasitology. 2003; 89(6):1250–4. https://doi.org/10.1645/GE-2786RN PMID: 14740922

12. Vega R, Pinero D, Ramanankandrasana B, Dumas M, Bouteille B, Fleury A, et al. Population genetic structure of *Taenia solium* from Madagascar and Mexico: implications for clinical profile diversity and immunological technology. International journal for parasitology. 2003; 33(13):1479–85. PMID: 14572511

13. Bobes RJ, Fragoso G, Reyes-Montes Mdel R, Duarte-Escalante E, Vega R, de Aluja AS, et al. Genetic diversity of *Taenia solium* cysticerci from naturally infected pigs of central Mexico. Veterinary parasitology. 2010; 168(1–2):130–5. https://doi.org/10.1016/j.vetpar.2009.11.001 PMID: 19963321

14. Maravilla P, Gonzalez-Guzman R, Zuniga G, Peniche A, Dominguez-Alpizar JL, Reyes-Montes R, et al. Genetic polymorphism in *Taenia solium* cysticerci recovered from experimental infections in pigs. Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2008; 8(2):213–6. https://doi.org/10.1016/j.meegid.2007.11.006 PMID: 18243817

15. Knapp J, Guislain MH, Bart JM, Raoul F, Gottstein B, Giraudoux P, et al. Genetic diversity of *Echinococcus multilocularis* on a local scale. Infect Genet Evol. 2008; 8(3):367–73. Epub 2008/04/15. https://doi.org/10.1016/j.meegid.2008.02.010 PMID: 18406214.

16. Casulli A, Bart JM, Knapp J, La Rosa G, Dusher G, Gottstein B, et al. Multi-locus microsatellite analysis supports the hypothesis of an autochthonous focus of *Echinococcus multilocularis* in northern Italy. Int J Parasitol. 2009; 39(7):837–42. Epub 2009/01/20. https://doi.org/10.1016/j.ijpara.2008.12.001 PMID: 19150351.

17. Casulli A, Szell Z, Pozio E, Sreter T. Spatial distribution and genetic diversity of *Echinococcus multilocularis* in Hungary. Vet Parasitol. 2010; 174(3–4):241–6. Epub 2010/10/01. https://doi.org/10.1016/j.vetpar.2010.08.023 PMID: 20880633.

18. Knapp J, Staebler S, Bart JM, Stien A, Yoccoz NG, Drogemuller C, et al. *Echinococcus multilocularis* in Svalbard, Norway: microsatellite genotyping to investigate the origin of a highly focal contamination. Infect Genet Evol. 2012; 12(6):1270–4. Epub 2012/04/03. https://doi.org/10.1016/j.meegid.2012.03.008 PMID: 22465539.

19. Knapp J, Bart JM, Maillard S, Gottstein B, Piarroux R. The genomic *Echinococcus microsatellite* EmsB sequences: from a molecular marker to the epidemiological tool. Parasitology. 2010; 137(3):439–49. Epub 2009/12/23. https://doi.org/10.1017/S0031182009991612 PMID: 20025824.

20. Umhang G, Knapp J, Hormaz V, Raoul F, Boue F. Using the genetics of *Echinococcus multilocularis* to trace the history of expansion from an endemic area. Infect Genet Evol. 2014; 22:142–9. Epub 2014/01/29. https://doi.org/10.1016/j.meegid.2014.01.018 PMID: 24468327.

21. Sprehn CG, Blum MJ, Quinn TP, Heins DC. Landscape genetics of Schistocephalus solidus parasites in threespine stickleback (*Gasterosteus aculeatus*) from Alaska. PLoS One. 2015; 10(4):e0122307. Epub 2015/04/16. https://doi.org/10.1371/journal.pone.0122307 PMID: 25874710; PubMed Central PMCID: PMCPMC4395347.

22. Detwiler JT, Criscione CD. Testing Mendelian inheritance from field-collected parasites: Revealing duplicated loci enables correct inference of reproductive mode and mating system. International journal for parasitology. 2011; 41(11):1185–95. https://doi.org/10.1016/j.ijpara.2011.07.003 PMID: 21839081

23. Pawlowski ZS. *Taenia solium*: Basic Biology and Transmission. In: Singh G, Prabhakar S, editors. Taenia solium Cysticercosis From Basic to Clinical Science. First. London, UK: CAB International; 2002. p. 1–13.

24. Tsang VC, Brand JA, Boyer AE. An enzyme-linked immunoelectrotransfer blot assay and glycoprotein antigens for diagnosing human cysticercosis (*Taenia solium*). The Journal of infectious diseases. 1989; 159(1):50–9. PMID: 2909643

25. Verastegui M, Gilman RH, Arana Y, Barber D, Velasquez J, Farfan M, et al. *Taenia solium* oncosphere adhesion to intestinal epithelial and Chinese hamster ovary cells in vitro. Infection and immunity. 2007; 75(11):5158–66. https://doi.org/10.1128/IAI.01175-06 PMID: 17698575

26. Deckers N, Kanobana K, Silva M, Gonzalez AE, Garcia HH, Gilman RH, et al. Serological responses in porcine cysticercosis: a link with the parasitological outcome of infection. International journal for parasitology. 2008; 38(10):1191–8. https://doi.org/10.1016/j.ijpara.2008.01.005 PMID: 18328486

27. de Aluja AS, Martinez MJJ, Villalobos AN. *Taenia solium* cysticercosis in young pigs: age at first infection and histological characteristics. Veterinary parasitology. 1998; 76(1–2):71–9. PMID: 9653992

28. Desjardins P, Conklin D. NanoDrop Microvolume Quantitation of Nucleic Acids. Journal of Visualized Experiments: JoVE. 2010;(45). https://doi.org/10.3791/2565 PMID: 21189466; PubMed Central PMCID: PMCPMC3346308.

29. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007; 23(10):1289–91. Epub 2007/03/24. https://doi.org/10.1093/bioinformatics/btm091 PMID: 17379693.

30. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012; 40(15):e115. Epub 2012/06/26. https://doi.org/10.1093/nar/gks596 PMID: 22730293; PubMed Central PMCID: PMCPMC3424584.

31. Watts NS, Pajuelo M, Clark T, Loader MC, Verastegui MR, Sterling C, et al. *Taenia solium* infection in Peru: a collaboration between Peace Corps Volunteers and researchers in a community based study. PloS one. 2014; 9(12):e113239. https://doi.org/10.1371/journal.pone.0113239 PMID: 25469506

32. Gonzalez AE, Cama V, Gilman RH, Tsang VC, Pilcher JB, Chavera A, et al. Prevalence and comparison of serologic assays, necropsy, and tongue examination for the diagnosis of porcine cysticercosis in Peru. The American Journal of Tropical Medicine and Hygiene. 1990; 43(2):194–9. PMID: 2389823

33. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010; 10(3):564–7. https://doi.org/10.1111/j.1755-0998.2010.02847.x PMID: 21565059.

34. Langella O. POPULATIONS 1.2.29. Population genetic software (individuals or populations distances, phylogenetic trees). 2002.

35. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer research. 1967; 27(2):209–20. Epub 1967/02/01. PMID: 6018555.

36. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

37. Sciutto E, Fragoso G, Fleury A, Laclette JP, Sotelo J, Aluja A, et al. *Taenia solium* disease in humans and pigs: an ancient parasitosis disease rooted in developing countries and emerging as a major health problem of global dimensions. Microbes and infection / Institut Pasteur. 2000; 2(15):1875–90.

38. Schlotterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic acids research. 1992; 20 (2):211–5. PMID: 1741246

39. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular ecology. 2002; 11(12):2453–65. PMID: 12453231

40. Thomas LF, de Glanville WA, Cook EA, Fevre EM. The spatial ecology of free-ranging domestic pigs (*Sus scrofa*) in western Kenya. BMC veterinary research. 2013; 9:46. Epub 2013/03/19. https://doi.org/10.1186/1746-6148-9-46 PMID: 23497587; PubMed Central PMCID: PMCPMC3637381.

41. Pray IW, Swanson DJ, Ayvar V, Muro C, Moyano LM, Gonzalez AE, et al. GPS Tracking of Free-Ranging Pigs to Evaluate Ring Strategies for the Control of Cysticercosis/Taeniasis in Peru. PLoS Negl Trop Dis. 2016; 10(4):e0004591. Epub 2016/04/02. https://doi.org/10.1371/journal.pntd.0004591 PMID: 27035825; PubMed Central PMCID: PMCPMC4818035.

42. Robinson DP, Klein SL. Pregnancy and pregnancy-associated hormones alter immune responses and disease pathogenesis. Hormones and behavior. 2012; 62(3):263–71. Epub 2012/03/13. https://doi.org/10.1016/j.yhbeh.2012.02.023 PMID: 22406114; PubMed Central PMCID: PMCPMC3376705.

43. de Aluja AS, Villalobos AN, Plancarte A, Rodarte LF, Hernandez M, Zamora C, et al. *Taenia solium* cysticercosis: immunity in pigs induced by primary infection. Vet Parasitol. 1999; 81(2):129–35. Epub 1999/02/25. PMID: 10030755.

44. Manrique P, Hoshi M, Fasabi M, Nolasco O, Yori P, Calderon M, et al. Assessment of an automated capillary system for *Plasmodium vivax* microsatellite genotyping. Malar J. 2015; 14:326. Epub 2015/08/22. https://doi.org/10.1186/s12936-015-0842-9 PMID: 26293655; PubMed Central PMCID: PMCPMC4546211.

PLOS | NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

Identification and Characterization of Microsatellite Markers Derived from the Whole Genome Analysis of *Taenia solium*

Mónica J. Pajuelo[1], María Eguiluz[1], Eric Dahlstrom[2], David Requena[1], Frank Guzmán[1], Manuel Ramirez[1], Patricia Sheen[1], Michael Frace[3], Scott Sammons[3], Vitaliano Cama[4], Sarah Anzick[2], Dan Bruno[2], Siddhartha Mahanty[2], Patricia Wilkins[4], Theodore Nash[2], Armando Gonzalez[5], Héctor H. García[6,7], Robert H. Gilman[8], Steve Porcella[2], Mirko Zimic[1]*, Cysticercosis Working Group in Peru[¶]

1 Laboratorio de Bioinformatica y Biologia Molecular, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru, **2** Genomics Unit, Research Technologies Section, Rocky Mountain Laboratories, NIAID, NIH, Hamilton, Montana, United States of America, **3** Biotechnology Core Facility Branch, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, **4** Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, **5** Facultad de Medicina Veterinaria, Universidad Nacional Mayor de San Marcos, Lima, Peru, **6** Departamento de Microbiología, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima Peru, **7** Instituto Nacional de Ciencias Neurológicas. Lima, Peru, **8** Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America

¶ Membership of the Cysticercosis Working Group in Peru is provided in the Acknowledgments.
* Mirko.zimic@upch.pe

OPEN ACCESS

Citation: Pajuelo MJ, Eguiluz M, Dahlstrom E, Requena D, Guzmán F, Ramirez M, et al. (2015) Identification and Characterization of Microsatellite Markers Derived from the Whole Genome Analysis of *Taenia solium*. PLoS Negl Trop Dis 9(12): e0004316. doi:10.1371/journal.pntd.0004316

Editor: Klaus Brehm, University of Würzburg, GERMANY

Received: August 31, 2015

Accepted: November 24, 2015

Published: December 23, 2015

Data Availability Statement: The set of contigs corresponding to the genome of the Huancayo-cysts, the set of contigs corresponding to the assembled genome of the Puno-proglottid and the set of contigs corresponding to the hybrid assembly created from both the Huancayo and Puno tissue genome assemblies is available together at Genbank under the project ID PRJNA183343.

Funding: This work was partially supported by the Fogarty International Center/NIH Training Grants D43TW001140 and D43TW006581. HHG. is supported by a Wellcome Trust Senior International

Abstract

Background

Infections with *Taenia solium* are the most common cause of adult acquired seizures worldwide, and are the leading cause of epilepsy in developing countries. A better understanding of the genetic diversity of *T. solium* will improve parasite diagnostics and transmission pathways in endemic areas thereby facilitating the design of future control measures and interventions. Microsatellite markers are useful genome features, which enable strain typing and identification in complex pathogen genomes. Here we describe microsatellite identification and characterization in *T. solium*, providing information that will assist in global efforts to control this important pathogen.

Methods

For genome sequencing, *T. solium* cysts and proglottids were collected from Huancayo and Puno in Peru, respectively. Using next generation sequencing (NGS) and *de novo* assembly, we assembled two draft genomes and one hybrid genome. Microsatellite sequences were identified and 36 of them were selected for further analysis. Twenty *T. solium* isolates were collected from Tumbes in the northern region, and twenty from Puno in the southern region of Peru. The size-polymorphism of the selected microsatellites was determined with

Research Fellowship in Public Health and Tropical Medicine. The funders had no role in study design, data collection and analysis or interpretaion; in writing the report or in the decision to submit this manuscript for publication.

Competing Interests: The authors have declared that no competing interests exist.

multi-capillary electrophoresis. We analyzed the association between microsatellite polymorphism and the geographic origin of the samples.

Results

The predicted size of the hybrid (proglottid genome combined with cyst genome) *T*. *solium* genome was 111 MB with a GC content of 42.54%. A total of 7,979 contigs (>1,000 nt) were obtained. We identified 9,129 microsatellites in the Puno-proglottid genome and 9,936 in the Huancayo-cyst genome, with 5 or more repeats, ranging from mono- to hexa-nucleotide. Seven microsatellites were polymorphic and 29 were monomorphic within the analyzed isolates. *T*. *solium* tapeworms were classified into two genetic groups that correlated with the North/South geographic origin of the parasites.

Conclusions/Significance

The availability of draft genomes for *T*. *solium* represents a significant step towards the understanding the biology of the parasite. We report here a set of *T*. *solium* polymorphic microsatellite markers that appear promising for genetic epidemiology studies.

Author Summary

Taenia solium, the pork tapeworm, is an important pathogen as it is a major cause of acquired epilepsy in developing countries. The parasite was eliminated from most developed countries decades ago due to improvement in sanitary conditions but it remains a common infection across Asia, Africa and Latin America. Identification of genetic variants within *T*. *solium* will enable to study the genetic epidemiology, distribution and movement of this parasite within endemic communities, which will ultimately facilitate the design of control strategies to reduce the health and economic burden of disease.

Microsatellites have been used in other parasites to identify genetic variants. In this study, we partially sequenced the genome of *T*. *solium* and identified microsatellites widely distributed in the genome using bioinformatics tools. We evaluated the distribution of these microsatellites collected from 20 tapeworms from the north and 20 tapeworms from the south of Peru. We identified seven polymorphic microsatellites, and evaluated their capacity to differentiate genetic variants of *T*. *solium*. Interestingly, tapeworms from the North and South of Peru showed different genotypes, suggesting its use as a potential marker to differentiate geographic origin.

Introduction

Cysticercosis is an infection caused by the larval stage of the cestode *Taenia solium*. When the larval stages infect the central nervous system, the infection is known as neurocysticercosis (NCC) and is the most common cause of adult-onset seizures in endemic regions worldwide. Crude estimates of the burden of infection and disease suggest that greater than ten million people have NCC and as many as 2.7–5.6 million suffer from epilepsy [1]. A recent analysis concluded that in Latin America, vast parts of Asia, the Indian subcontinent and Southern China, Sub-Saharan Africa, and Oceania, 29% of all cases of epilepsy are attributable to NCC [2].

Humans are the only known definitive host, harboring the adult tapeworm and releasing infectious eggs to the environment [3]. In pigs that ingest infectious ova or proglottids, the released oncospheres cross the intestinal wall into the circulatory system where they become trapped in the microcapilaries, often in the brain, muscles and subcutaneous tissues. The oncospheres develop into cysticerci (cysts) and if present in the parenchyma of the brain, seizures and epilepsy may occur as a result of host inflammation against the cysts.

Understanding the genetic variation of *T. solium* has the potential to improve our knowledge of the biology, epidemiology, infectivity, and pathogenicity of this parasite in endemic regions [4–7]. Moreover, analysis of the genetic variation within and between different geographical populations can provide information on evolution [8], genetic differentiation and speciation of parasites [9], as well as provide tools for understanding transmission dynamics, which may contribute to public health efforts to control this parasitic infection.

The first attempts at genotyping *Taenia* parasites were directed towards the differentiation of *Taenia* species based on the sequence polymorphism of mitochondrial NADH dehydrogenase 1 and cytochrome c oxidase subunit I (COI) genes using single-strand conformation polymorphism (SSCP) [10]. Restriction fragment length polymorphism (RFLP) also was used to discriminate *Taenia* species by analyzing the ribosomal 5.8S gene sequence as well as the internal transcribed spacer (ITS) [11]. In 2001, Hancock *et al* showed diversity among *T. solium* cysts from different countries using COI, a portion of the ITS1 encoded gene and the diagnostic antigen Ts14. Little genetic diversity within *T. solium* samples collected from South America and Asia was observed. In addition, 15 isolates from Peru had similar COI sequences showing no genetic variability between them [12]. Later, two different worldwide genotypes were reported, with Asian parasites grouped into one cluster, and parasites from Latin America and Africa grouped into another cluster [4]. It has been suggested that the low variation found in *T. solium* isolates may be associated with the limited resolution of the experimental techniques used at that time [8].

More recently, with the development of new DNA analysis tools such as Random Amplification of Polymorphic DNA (RAPD), greater genetic variation has been reported in parasites from communities in Mexico, Honduras and Madagascar [5,13–15]. This data suggests that *T. solium* has local lineages with different genetic characteristics. However, some disadvantages have been reported with RAPD such as low reproducibility, inability to test heterozygosity and subjective interpretation of the data [16,17]. Therefore, a more robust tool with higher resolution is needed to obtain more precise genotyping of *T. solium* isolates.

Microsatellites, or Simple Sequence Repeats (SSR), are repetitive DNA sequences consisting of blocks of 1 to 6 nucleotides repeated up to 60 times 8. They are highly polymorphic in the number of repeated units. The variation in size of repeat domains is mainly generated by slippage of DNA polymerase during DNA replication, resulting in the insertion or deletion of one or more repeated units [18,19]. Microsatellites have the advantage of being multi-loci and principally neutral markers [19], meaning that unlike protein-encoding genes, they are less likely to be subject to selective pressure. Microsatellites are highly reproducible and specific, and are easily identified from genome sequences by bioinformatics data mining [20–22].

Microsatellite polymorphisms can be detected by polymerase chain reaction (PCR) amplification followed by DNA electrophoresis [8,23]. This technique has been used to analyze genetic variation of other parasites such as *Leishmania* spp. [24], *Schistosoma japonicum* [6], *Trypanosoma cruzi* [25], and *Plasmodium falciparum* [26]. Microsatellite markers also have the advantage of being able to detect greater genetic variation than other genetic markers, as has been demonstrated in *Echinococcus multilocularis* [27]. This work suggests that microsatellite polymorphism analysis is an appropriate tool to differentiate *T. solium* isolates. With the availability of draft genome sequences, the identification of microsatellites is more efficient [20].

Recently, a draft genome of a *T. solium* isolate recovered from Mexico has been published [28]. It is however necessary to have more genomic information available, in order to identify genotyping markers.

In this study we present two draft genome sequences corresponding to *T. solium* specimens from Huancayo (cysts) and Puno (proglottid) from which we identified and characterized microsatellite markers. We explore microsatellite length variability to differentiate *T. solium* isolates from two regions of Peru. To analyze the microsatellites length we used a multi-capillary electrophoresis QIAxcel system that has the advantage of automatized size determination [29]. Although its advantages, this technique is limited by 3–5bp resolution that will not let us differentiate length polymorphism lower than 3–5 bp. The proposed microsatellites will lead to a more comprehensive understanding of the epidemiology of this important human pathogen.

Materials and Methods

Taenia solium tissue specimens for genome sequencing

Individual cysticerci (cysts) were recovered from a single, naturally infected pig from Huancayo, a city in the central Andean region. One proglottid from Puno (Fig 1) was excised from segments of an adult tapeworm recovered from a single fecal specimen and used for extraction of DNA. To minimize contamination with exogenous materials, the proglottid was washed thoroughly 10 times with phosphate buffer solution (PBS), transported in a mixture of PBS and antibiotics penicillin/streptomycin/ amphotericin B), and stored at -70°C until the DNA was extracted.

Extraction of genomic DNAs from Huancayo cysts and Puno proglottids

DNA extraction from the Huancayo cyst sample. DNA and RNA was extracted simultaneously from *T. solium* Huancayo cysts (named "Huancayo-cyst") at Rocky Mountain

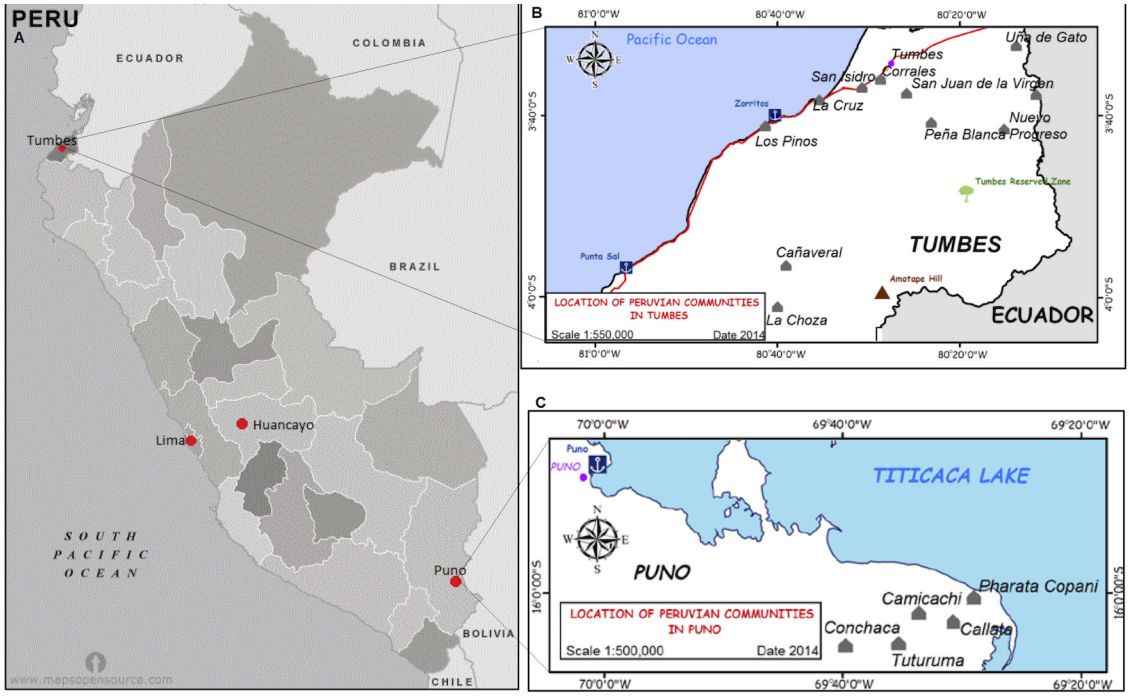


Fig 1. Map of Peru depicting the location of *T. solium* isolates used for this study. Map showing the location of Tumbes and Puno, as well as Huancayo (locality of one of the genome source) and Lima, the country´s capital (A). Maps of Tumbes (B) and Puno (C) showing the location of the communities of origin of the tapeworms for this study

doi:10.1371/journal.pntd.0004316.g001

Laboratories (RML) (NIH, NIAID) following the Qiagen AllPrep DNA/RNA Mini kit (Qiagen, Valencia, CA) protocol with the following modifications. Cysts were first frozen in liquid nitrogen and finely ground to a powdery-like texture using a liquid nitrogen chilled mortar and pestle. The ground tissue was transferred to a 1.5 mL Eppendorf tube containing 1.0 mL RLT buffer (cell lysis buffer for DNA/RNA extraction, Qiagen) supplemented with 0.143 M β-mercaptoethanol and allowed to incubate at room temperature for 15 minutes with occasional gentle vortexing. Following the incubation, the extract was homogenized using a QIAshredder spin column (Qiagen) and then purified with AllPrep DNA/RNA kit reagents following the manufacturer's protocol. DNA and RNAs were both eluted twice with 50 μl elution buffer per elution. DNA yield and purity was assessed using dsDNA quantitation (Life Technologies, Grand Island, NY) and UV spectrophotometry at A260 nm and A280, respectively. DNA quality was visualized on a 1% agarose gel (Lonza, Rockland, ME). Portions of the DNA samples from the cysts were all pooled into one tube in order to meet quantity and quality thresholds required for downstream Next Generation Illumina sequencing. RNAs were stored at -80°C for future mRNA analyses.

DNA extraction from the Puno proglottid sample. DNA was extracted from a *T. solium* proglottid originating from Puno (named "Puno-proglottid") at the Centers for Disease Control and Prevention (CDC). Genomic DNA was extracted using a phenol chloroform extraction method, which included a lysozyme treatment at 37°C for 1 hour to minimize/eliminate bacterial contamination of proglottids, since those were recovered from human feces, followed by incubation with Proteinase K at 56°C overnight. The quantity and quality of the eluted DNA was initially tested by ultraviolet-Vis spectrophotometry (Nanodrop, Thermo Scientific, Newark, DE), and further evaluated with a picogreen kit (Life Technologies, Cat N°P7598) and Biophotometer Plus (Eppendorf), respectively. In addition, for quality analysis, one microliter of the eluted DNA was evaluated by electrophoresis through a 2% agarose gel.

Next Generation library construction and sequencing procedure

Sequencing of the Huancayo-cysts. For the Huancayo-cyst DNA, a paired-end and mate-pair library were constructed separately and run on an Illumina HiSeq 2000 (Illumina Inc, San Diego CA) at RML/NIH. The paired-end library was prepared as follows; one microgram of pooled Huancayo cyst DNA was processed using the TruSeq DNA Sample Preparation Guide, Rev. A., November 2010 (Illumina Inc., San Diego, CA). The resulting library was clustered on a cBot using a paired-end flowcell and sequenced 100 cycles (bases) from both ends. The mate-pair library was prepared and sequenced in the following manner; approximately 4 micrograms of pooled Huancayo-cyst DNA was sheared using a Hydroshear device (Digilab Inc., Marlborough, MA) and processed following the procedure provided in Preparing 2–5kb Samples for Mate Pair Library Sequencing, Rev. B, February 2009 (Illumina Inc., San Diego, CA). The biotin-labeled fragments were electrophoresed on a 0.8% TAE preparatory gel and fragments ranging in size from 2.4 to 4 Kb were excised for the circularization reaction. The resulting library was clustered on a cBot using a paired-end flowcell and sequenced 100 cycles (bases) from both ends.

Sequencing of the Puno-proglottid. For the Puno-proglottid DNA a fragment sequencing approach using the 454 GS-FLX+ (Roche Applied Science, Indianapolis, IN) and Illumina Genome Analyzer IIe (Illumina Inc., San Diego, CA) was employed at the CDC. The 454-FLX Titanium shotgun library was prepared as follows; DNA fragments larger than 1.5 Kb were extracted from the Puno proglottid genomic DNA and 500 ng was processed using the procedure in the Rapid Library Preparation Method Manual, GS FLX+ Titanium Series, October 2009 (454 Life Sciences, Branford, CT). For emulsion PCR a 4 copy-per-bead ratio was used to

yield a bead enrichment of 8%. The resulting enriched beads were then sequenced using 454 GS FLX Titanium chemistry. The Illumina fragment library was prepared and sequenced in the following manner: one microgram of Puno proglottid genomic DNA was processed using TruSeq DNA Sample Preparation Guide, Rev. A., November 2010 (Illumina Inc., San Diego, CA). The resulting library was clustered on a cBot using a single-read flowcell and sequenced 100 cycles (bases).

Huancayo-cyst genome assembly

For the assembly, both the Illumina mate-pair and paired-end sequencing results were combined, providing a total of 878,340,445 usable reads at 100 bp average length. In order to provide the best assembly at an average of 157X coverage and for an expected genome size of 115MB, a subset of 175,931,369 reads were selected for de novo assembly process using Velvet v1.1.05 [30], with the following parameters: coverage cutoff = 10, expected coverage = 26, paired end insert length = 350, and mate paired insert length = 3,500.

Puno-proglottid genome assembly

For the Puno proglottid genome assembly, both the Illumina mate-pair and the 454 GS FLX raw signals data were processed with the software GS Run Processor to obtain the reads, which were assembled using GS de novo Assembler 2.6 with the following parameters: minimum read length = 20 nucleotides (nt), overlap seed step = 12 nt, overlap seed length = 16 nt, overlap minimum match length = 40 nt, overlap minimum match identity = 90 nt, overlap match identity score = 2 nt, overlap match difference score = -3 nt, all contig threshold = 100 nt, large contig threshold = 500 nt; with an expected depth of 25X. Duplicated reads were used and the assembly was performed 30 times, taking the iteration with the mean number of contigs.

Cysticercus/proglottid hybrid genome assembly

To generate a more complete *T*. *solium* genome sequence, we produced a hybrid assembly consisting of the Puno-proglottid fragment 454 reads combined with the Huancayo-cyst Illumina paired-end reads as follows. The hybrid genome was generated using Velvet v1.1.05 de novo assembler using the Puno-proglottid 454 fragment reads, Huancayo-cyst Illumina paired-end reads, and Huancayo-cyst Illumina mate-pair reads. Velvet parameters used are as follows: kmer value = 57, coverage cut-off = 9, expected coverage = 38, paired insert length = 350, and mate pair insert length = 3500.

Public submission of genomes assemblies

The set of contigs corresponding to the genome of the Huancayo-cysts, the set of contigs corresponding to the assembled genome of the Puno-proglottid and the set of contigs corresponding to the hybrid assembly created from both the Huancayo and Puno tissue genome assemblies is available together at Genbank under the project ID PRJNA183343.

Identification of microsatellites in the *T*. *solium* genome

Microsatellites were identified using the script developed by Gur-Arie *et al* [31]. We searched for repetitive motifs of 1–6 bp in our two assembled *T*. *solium* genomes. A total of 36 distinct microsatellites with polymorphic attributes were selected according to the following criteria: For the Puno-proglottid genome, a first group of microsatellites with a minimum of five motif repetitions were selected. In order to determine if the microsatellites are transcribed, we verified if they were present in the *T*. *solium* ESTs sequences database available (http://www.ncbi.

nlm.nih.gov/nucest/?term=%22Taenia+solium%22%5Bporgn%3A__txid6204%5D), which later was included in the published *T. solium* genome sequence [28]. In order to determine if the microsatellites are comprised within a coding region, we verified the presence of an ORF using the algorithm ORF-Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Microsatellites that showed differences in the number of repeats between the Puno-proglottid genome and the ESTs database were selected for further analysis. We verified that conserved flanking regions were present in both the Puno-proglottid genome and the ESTs database. Five microsatellites (TS_SSR01 to T S_SSR05) were selected. In addition, 8 of the longest microsatellite sequences from the Puno-proglottid genome not present in the ESTs database also were selected (TS_SSR06 to TS_SSR13). Because this first group of microsatellites was biased based on their availability in ESTs, it is less likely that they are neutral.

A second group of microsatellites was identified in the Puno-proglottid genome and mapped into the corresponding contigs of the Huancayo cyst genome. 200 microsatellites (100 di-nucleotides and 100 tri-nucleotides) with the largest repetitive motif in both genomes were evaluated. After aligning the sequences we selected the largest sequences that showed polymorphisms in the repetitive sequences between both genomes (13–18 repeats). This resulted in additional 23 microsatellite loci for testing.

Primer design for microsatellite amplification. In order to prevent any bias due to the assembly of a consensus genome that does not account for the intrinsic genomic variability due to the diplodism of *T. solium* and the potentially different sources of infection giving raise to the multiple cysticercus processed to obtain DNA, we directly determined the size of the microsatellites after the individual amplification in each sample. Specific primers were designed against the conserved flanking sequences of each microsatellite. The Primer3 program was used with the following parameters: 20 bp length, 50% GC and 50°C Tm [32].

Tapeworm specimens. A total of 40 *T. solium* adult tapeworm specimens were randomly and anonymously selected from the repository of the Cysticercosis Working Group in Peru. Isolates came from different communities from the northern city of Peru, Tumbes (N = 20) and the southern city of Peru, Puno (N = 20) (Table 1). The geographical distribution of the collected *T. solium* tapeworm isolates is shown in Fig 1. Tapeworms were stored in 25% glycerol supplemented with penicillin (1000 UI/mL) and Gentamycin (100 μg/mL) at 4°C until use.

DNA purification and PCR amplification of microsatellite loci. DNA purification was performed using the QIAmp DNA Mini Kit (QIAGEN) according to manufacturer´s instructions. DNA was quantified by UV spectrophotometry (Nanodrop) and stored at -20°C until use. The PCR reaction volume (25 μL) consisted of: Buffer 1X (Promega), 2.5 mM $MgCl_2$, 0.2 mM dNTPs each one, Forward primer: 1μmol, Reverse primer: 1 μmol and 20 ng of DNA, Taq polymerase 1 U. PCR was conducted in a MJ Research MiniCycler PTC-150 thermocycler with a hot cover using the following temperature profile except where otherwise stated: the initial denaturation step was at 94°C for 4 minutes, followed by 30 cycles of 94°C for 1 minute, 55–65°C for 1 minute and 72°C for 1 minute and a final extension at 72°C for 10 minutes. Microsatellite markers were analyzed in a multi-capillary electrophoresis QIAxcel system using the QIAxcel high-resolution kit with the OM700 method [29]. 25 bp-500 bp ladder was used as a size marker for the assignment of the allele sizes using the system's software. As a control and in order to verify reproducibility, we amplified the microsatellite TS_SSR01 in four independent replicates from 9 DNA samples.

Association analysis of microsatellite polymorphism with geographic location. In order to determine an association between the geographic origins of the tapeworms with the specific microsatellites, we performed a Fischer's exact test to evaluate the independence between the alleles identified in each microsatellite and the north/south geographic origin.

Table 1. Origin of *T. solium* tapeworm isolates for the evaluation of microsatellites.

Tapeworm ID	Community	Populations †	Region
6	San Isidro	Cruz-Isidro-pinos	Tumbes
39	San Isidro	Cruz-Isidro-pinos	Tumbes
28	San Isidro	Cruz-Isidro-pinos	Tumbes
38	La Cruz	Cruz-Isidro-pinos	Tumbes
7	Los Pinos	Cruz-Isidro-Pinos	Tumbes
19	Tumbes	Tum-Corrales-San Juan-PenaBlanca	Tumbes
31	Tumbes	Tum_corrales-san juan_PenaBlanca	Tumbes
20	Corrales	Tum_corrales-san juan_PenaBlanca	Tumbes
27	Corrales	Tum_corrales-san juan_PenaBlanca	Tumbes
11	Corrales	Tum_corrales-san juan_PenaBlanca	Tumbes
12	Fuerte 5 de Julio	Tum_corrales-san juan_PenaBlanca	Tumbes
5	San Juan—Virgen	Tum_corrales-san juan_PenaBlanca	Tumbes
25	Peña Blanca	Tum_corrales-san juan_PenaBlanca	Tumbes
18	La Choza	Cañaveral–Choza	Tumbes
24	Cañaveral	Cañaveral–Choza	Tumbes
14	Pueblo Nuevo	Uñagato-progreso-pueblonuevo	Tumbes
36	Pueblo Nuevo	Uñagato-progreso-pueblonuevo	Tumbes
8	Nuevo Progreso	Uñagato-progreso-pueblonuevo	Tumbes
35	Nuevo Progreso	Uñagato-progreso-pueblonuevo	Tumbes
13	Uña de Gato	Uñagato-progreso-pueblonuevo	Tumbes
40	Pharata	Fharata	Puno
1	Pharata	Fharata	Puno
30	Pharata	Fharata	Puno
17	Pharata	Fharata	Puno
21	Pharata	Fharata	Puno
34	Pharata	Fharata	Puno
3	Callata	Callata	Puno
9	Callata	Callata	Puno
10	Callata	Callata	Puno
29	Callata	Callata	Puno
4	Callata	Callata	Puno
15	Camicachi	Camicachi	Puno
33	Camicachi	Camicachi	Puno
37	Camicachi	Camicachi	Puno
23	Camicachi	Camicachi	Puno
32	Camicachi	Camicachi	Puno
16	Conchaca	Conchaca_Tuturuma	Puno
2	Conchaca	Conchaca_Tuturuma	Puno
26	Conchaca	Conchaca_Tuturuma	Puno
22	Tuturuma	Conchaca_Tuturuma	Puno

† Populations are the groups that were formed by proximity for the phylogenetic analysis

doi:10.1371/journal.pntd.0004316.t001

Ethics statement

Taenia solium cysts were excided in a previous study from a naturally infected pig in a Huancayo local abattoir. The pig was bought by the study team at market price so the study team owned the animal. Procedures were approved by Universidad Peruana Cayetano Heredia

(UPCH) ethics committee for animal use *T. solium* proglottid specimens were collected in previous studies by the Cysticercosis Working Group in Peru, with approval of the UPCH IRB (IRB00001014); they were used as residual diagnostic samples.

Results

Analysis of genomic assemblies

Sequencing of the Huancayo-cyst genome produced a total of 175,931,369 reads that were assembled into 18,361 contigs (the largest contig was 307,365 bp). The estimated genome size was 114,605,177 nt. The sequencing of the Puno-proglottid genome produced a total of 76,625,473 reads that were assembled in 47,475 contigs (the largest contig was 79,438 nt). The lack of large contigs in this assembly is due to the lack of paired-end sequencing data. The estimated size of the Puno-proglottid genome was 109,898,809 nt. For the hybrid cyst/proglottid assembly, 7,979 contigs were obtained, and the largest contig had 395,362 nt. The estimated size of the hybrid genome was 111,029,218 nt. These and other statistics were calculated with the program multifastats,py v1.4 (https://github.com/lbbm-upch/multifastats_v1.4), and are summarized in Table 2. As expected due to their closeness, the Peruvian *T. solium* genomes showed a similar size as the recently published Mexican *T. solium* genome (122.3 Mb), as well as the genomes of *E. granulosus* (114.9 Mb), and *E. multilocularis* (115 Mb) [28].

Identification of microsatellites

We identified 9,129 microsatellite sequences distributed in the *T. solium* Puno-proglottid genome and 9,936 in the Huancayo-cyst Genome (Table 3). In both the Puno and Huancayo genomes, the greatest number of microsatellite loci found contained di-nucleotides repeats. Most of these microsatellites were over-represented in the forms of AC/GT and AG/CT, while the forms AT/AT and CG/CG showed a lower frequency of occurrence (S1 Table).

Microsatellite PCR and polymorphism analysis

Thirty-six microsatellites markers were identified as potentially polymorphic. We successfully amplified 34 microsatellite markers in 40 *T. solium* tapeworm specimens. The estimated size of the PCR products of the microsatellites in the tapeworm isolates was similar to the expected theoretical sizes.

Table 2. *Taenia solium* genomes assemblies' statistics.

	CDC	NIH	Hybrid	Hybrid
	Puno proglottid	Huancayo Cyst	Cyst/proglottid (all contigs)	Cyst/proglottid (GenBank)†
Total sequenced bases	109,898,809	114,605,177	116,668,703	111,029,218
Contigs	47,475	18,361	19,727	7,979
GC%	42.96	42.83	42.86	42.80
N50	4,839	39,744	46,836	43,923
Shortest contig (nt)	287	51	51	1000
Largest contig (nt)	79,438	307,365	395,362	39,5362
Mean contig length	2,314.9	6,241.8	5,914.2	13,915.2

Statistics of the hybrid genome are shown for the full version and for the set of contigs uploaded to GenBank
†Only contigs larger than 1000bp were submitted

doi:10.1371/journal.pntd.0004316.t002

Table 3. Frequency of microsatellites per type found in *T. solium* partial Genome 1 and partial Genome 2.

Number of nucleotides	Range of repeats	Number of microsatellites	
		Genome 1	Genome 2
Total bp analized		109'898,809	114,605,177
Mono nucleotide	10 to 53	2,298	2,737
Di nucleotide	6 to 38	3,393	3,537
Tri nucleotide	5 to 28	2,393	3,464
Tetra nucleotide	5 to 28	889	1,044
Penta nucleotide	5 to 18	123	122
Hexa nucleotide	5 to 10	33	32
Total		9,129	9,936

doi:10.1371/journal.pntd.0004316.t003

Within the tested sample, twenty-seven microsatellite markers were monomorphic, of which 26 were homozygous (only one band was observed in the electrophoretic pattern) and 1 was heterozygous (TS_SSR31, in which two bands were observed in all samples). Seven microsatellite markers were polymorphic containing a total of 44 alleles within the polymorphic loci (Table 4). All 40 tapeworms were homozygous for five markers (TS_SSR09, TS_SSR16, TS_SSR18, TS_SSR27, and TS_SSR28), while some were heterozygous for TS_SSR01 and TS_SSR32 (S2 Table). The number of alleles varied from 4 for locus TS_SSR16 to 10 for locus TS_SSR32 with an average of 6 alleles per locus. The polymorphic information content (PIC) varied from 0.472 for locus TS_SSR16 to 0.843 for locus TS_SSR28 with an average of 0.604 per locus (Table 4).

The reproducibility analysis of TS_SSR01 amplification showed a variability of 1–2 bp between the four replicas in the nine isolates tested (S3 Table), which is lower than the range of resolution reported by the manufacturer (3–5 bp). The variability of TS_SSR01 size between the different *T. solium* isolates appeared as 2–15 bp, which is three fold higher than the experimental error reported by the manufacturer (S3 Table).

Only TS_SSR01 was found to be present in the EST database. However the comprising region did not show evidence of any ORF. Therefore TS_SSR01 is part of a transcribed but not-translated sequence. When we compared TS_SSR01 against the complete no-redundant nucleotide collection of Genbank using blastn, only one sequence from the close organism *T. asiatica* appeared similar.

Association of microsatellite polymorphism with the geographic origin of tapeworms

The genetic diversity observed in isolates from the southern city of Puno (20 different genotypes) was slightly higher than that the genetic diversity observed in the isolates from the northern city Tumbes (16 different genotypes). Also the median number of alleles was slightly higher (Table 5). The single microsatellite that best differentiated tapeworms from Tumbes and Puno isolates was TS_SSR01. Most of the isolates from Tumbes (17/20) were associated to genotype A (206/206 bp) and 3/20 of isolates were associated to genotype B (211/211 bp), all of them homozygous. A lower prevalence of genotype A was observed in Puno (6/20) compared to Tumbes (P = 0.001, Fisher's exact test). A similar prevalence of genotype B (2/20) was observed in Puno. Five other genotypes were found only in Puno (S2 Table).

Discussion

The present study describes the draft genome sequences of two *T. solium* isolates and the identification and characterization of DNA microsatellites. The *T. solium* microsatellites reported

Table 4. Characteristics of microsatellite loci evaluated.

Microsatellite	Primer sequences (5'-3')	Repetition motif	Observed size (bp)	Number of Alleles	PIC ±
TS_SSR_01	ACCGGTGGTCGGAATTATTA GTTCTTGCTGAGGTGGTTCC	(CCATT)	206–226	5	0.582
TS_SSR_02	CTCCGTGTCTTGACAGCAAA TGACGAAATGGAACAGTGGA	(ATGA)	190	1	-
TS_SSR_03	TTTCAAGCACGTGTCAGCAT GCTGGCAGACAGTGAGTAGG	(CATT)	155	1	-
TS_SSR_04	CAGATGAGGGGATGATGCTT GAACGATCCCAACCTCCATA	(GTT)	180	1	-
TS_SSR_05	GGGAAAATGCAGTTCAGAGC GGTCTGATGCGAGGTCTAGG	(TAA)	197	1	-
TS_SSR_06	GACCAAGCCCAACACCTCTA CAAGAATGAACGGGAGCAAC	(GGTA)	177	1	-
TS_SSR_07	GCACACAAACTGGTCACTCG TGCTATGCGTTTGCTTGTTT	(CAAT)	–	-	-
TS_SSR_08	TCGTCAGTGTGGGAGAGTGA TGGTTGGATTTGTGCTTTGA	(ACG)	–	-	-
TS_SSR_09	AAGCCAATGGTGACCAAGAG GCCAGCATAGAAGAGCCTGT	(GGT)	166–178	5	0.534
TS_SSR_10	CGACTCACGGCATTCATCTA TCCAAGACCCTGTGAAATCC	(GT)	220	1	-
TS_SSR_11	TCATCTTCCCCGTAAGGCTA ACACTCGAAGCGCAGTGTTT	(GA)	181	1	-
TS_SSR_12	ATCTCGACAGGCTCGAGTTC TCCGAACAGCTTCGAGTTTT	(TG)	192	1	-
TS_SSR_13	GTAGCGGTAACGGAGTGAGG TCAGGCTGGTAACGTGTCAG	(GT)	202	1	-
TS_SSR_14	AGCCGGTTCTCAGTTGATTG AATGCACTCATGCCATCTCA	(TG)	162	1	-
TS_SSR_15	GAAAAGAACGGCATGCAAAT GTTTGGCCATTTTGCCTCTA	(AT)	165	1	-
TS_SSR_16	CGCTGGACTAGGGTCGAATA CAGCAGAACAACAGCACCAT	(GT)	160–166	4	0.472
TS_SSR_17	GCATTCCGAGGATGAATGAT CGTTTTTCTGCACACTTGGA	(CA)	160	1	-
TS_SSR_18	AGTTAGCGTGCTTGCTTGGT ATTCCTGTTGCAACCTCCAC	(GT)	168–180	6	0.638
TS_SSR_19	TCCCTTACACCCTTCACGTC AAAGGCGGTAGATTGTGTGC	(TG)	163	1	-
TS_SSR_20	GGCCATTCAGTACCAACCAT TGTGCATGCCATTGTATGTG	(CT)	154	1	-
TS_SSR_21	CTATGCCACACCCAACAATG GGCCTTCAAGATCACTCGTC	(GT)	187	1	-
TS_SSR_22	CCTATTCCACTGGGGTGATG TCGATGAGCTTGCTGTATGTG	(TG)	178	1	-
TS_SSR_23	CCTTTTTCGGTGAAGTCGAT GCCTCCTTACACACATGCAA	(CA)	209	1	-
TS_SSR_24	CCCCATTTCCTGTTTCCTCT GCGGTGGCAATATAAGCATT	(CT)	144	1	-
TS_SSR_25	AGGTGGCGTTATGAATCAGC GCAAACCATCGGATAAAGGA	(AC)	174	1	-
TS_SSR_26	CGGTTTGCTTTTATGCCAAT AAATGGTCGCCTGAAATGAC	(GAA)	165	1	-
TS_SSR_27	GAGGTCTCGCCTCATCAAAG TTTCCACTCCCAAAAACTCG	(GAA)	158–176	5	0.546
TS_SSR_28	TGACGCTGGTAAGCTGTTTG GGAACCTTGGCACGAGATAG	(GTA)	202–226	9	0.843
TS_SSR_29	AAAGATGGACGGAAACAGGA GTTGGACGGAGATGTGTGTG	(AGG)	187	1	-
TS_SSR_30	TGACGTGTCGTCAGGTAGGA CGCATAGCCAGTACTTGTTCC	(TCC)	190	1	-
TS_SSR_31	GGTTGCTTTTGCTTGTCCTC CACTCTCCACGAGTCCACAA	(TGA)	157/179	1	-
TS_SSR_32	TGACGTTAACGAGGGTGTTG AGATCTCGCCTTGCAACAAT	(AGC)	177–210	10	0.617
TS_SSR_33	CCAGCGGCATATTACAAAGG ACTCAAAAGCGCCGAAATTA	(AGG)	130	1	-
TS_SSR_34	ATCACTCCTGTCCCAACTGC GGGTCGATTGGTCAGAGAAA	(CCT)	182	1	-
TS_SSR_35	GGGCGTGAACTCGAATAAAA GGGGCAGACAAGTGAAAAAG	(CCA)	170	1	-
TS_SSR_36	GCCCTGATTGTTGCTTTTGT AACGACACGCGGAAAATATC	(TCT)	175	1	-

± PIC was calculated only for polymorphic microsatellites

doi:10.1371/journal.pntd.0004316.t004

Table 5. General characteristics of polymorphic microsatellites by region.

Region	Number of isolates	Number of polymorphic loci	Median number of alleles per locus	Total number of different genotypes
Tumbes	20	7	3	16
Puno	20	7	5	20

doi:10.1371/journal.pntd.0004316.t005

here were found to be distributed along the entire genome. The length polymorphism of microsatellites was analyzed for its association with the geographic origin of tapeworm isolates. We found novel microsatellites that were able to differentiate tapeworms between the northern and southern regions of Peru.

Microsatellites have proven to be highly informative in population genetic studies in several parasites [6,33]. In the particular case of *T. solium*, the use of microsatellite markers allows a way to define the genetic structure of populations and to conduct genetic epidemiology studies. Although previous studies have shown a moderate genetic diversity of *T. solium* [4,9,10,34] and particularly in Peru [12], the novel microsatellites we identified here have demonstrated the capacity to differentiate tapeworms from Tumbes in the north and Puno in the south of Peru.

Although the frequency of microsatellites and their coverage in the genome varies considerably between organisms, the number of microsatellites found in *T. solium* (between 9,000–10,000) is similar to the number of microsatellites identified in other parasites [21,35].

Although most of the microsatellite sequences were found in non-transcribed regions, we found that *T. solium* microsatellites could also be present in transcribed/non-translated regions, being the abundance in non-transcribing regions higher than in transcribed/non-translated regions. This result is consistent with previous studies that reported that microsatellites are more abundant in non-coding regions of eukaryotic organisms [36]. The relatively low abundance of microsatellites in transcribed/non-translated as well as in coding regions could be explained by a negative selection against mutations that change the function by altering the secondary structure of the transcribed sequence or by altering the reading frame of the genes [36,37].

Eukaryote microsatellite loci typically contain between 5 and 40 repeats, similar to what we found in *T. solium*. As in other organisms, the number of microsatellites in *T. solium* decreases as the size of the repeat unit increases [38,39]. It is important to highlight that the distribution of the repeat types (mono- to hexa-nucleotide) varies across different taxa, and it has been suggested that this variation is associated with to the interaction of the mutation and the differential selection pressure [37].

As previously reported in other species, we found that dinucleotide repeats motifs were the most abundant type in *T. solium*, which tend to be longer in non-coding regions. This seems to be explained by the negative selection pressure of polymerase slippage during replication of coding DNA [40]. Castagnone—Sereno *et al.* reported that in nematodes, (AT)n was the most common microsatellite motif [35]. We found the AC / GT dinucleotide motif to be the most abundant in *T. solium*, which concordantly has also been found to be common in most vertebrates and arthropods [41].

The genetic variability observed in this study may be explained by several factors, including migration of humans and pigs, mutations in the tapeworm genome, cross-fertilization of tapeworms in the intestine in cases where multiple tapeworm infections occur [42,43], among others.

It is important to note that although the low resolution of QIAxcel system reported by the company (3–5bp), the range of difference in the size of TS SSR01 between the north and south region of Peru, is 2–3 fold higher than the expected error (5–15 bp) and the results of the repeatability assay showed lower variability (1–2 bp). This evidence supports the main finding of having TS SSR01 as a polymorphic marker able to differentiate tapeworms from the north and south region of Peru.

Transmission dynamics are not fully understood, although genetic characterization by means of microsatellite genotyping may unveil details of the ecology of *T. solium*. The use of molecular characterization by means of microsatellites will potentially allow identification of

genetic links between tapeworms, larval cysts found in infected pigs and eggs in soil or fomites. Furthermore, microsatellites would help disentangle the genetic complexity of a population due to the introduction of external tapeworms from immigrant tapeworm-carriers. This method of genotyping also has implications in the evaluation of parasite control by identifying the source of infection, and the re-introduction routes of the parasite into a specific region.

In conclusion, this study describes the identification and application of microsatellite markers in *T. solium* genotyping. The novel microsatellites reported here would be an important tool for future studies of the genetic variability of *T. solium*, including population genetics, basic epidemiology, super infections with more than one strain, and tracking the transmission of cysticercosis.

Supporting Information

S1 Table. Relative frequency of different motifs in each type of microsatellites: mono-, di- and tri- nucleotides in the partial genomes of *Taenia solium*.
(DOCX)

S2 Table. Genotype of the 40 *Taenia solium* isolates for each polymorphic microsatellite marker.
(DOCX)

S3 Table. Repeatability assay of microsatellite TS_SSR01.
(DOCX)

Acknowledgments

We thank our UPCH students Katherine Lozano, Sebastian Carrasco, Basilio Cieza, Vladimir Espinoza, Eduardo Gushiken, and Bryan Lucero for their participation in the initial bioinformatics exploratory analysis. We are grateful to Dr. Seth O´Neal for his valuable and helpful comments regarding the manuscript and to Eng. Carmen Gamero Huamán for creating the maps of the communities presented in this article.

Members of the Cysticercosis Working Group in Peru

Victor C.W.Tsang, PhD (Coordination Board); Silvia Rodríguez, MSc; Isidro González, MD; Herbert Saavedra, MD; Manuel Martínez, MD; Manuel Alvarado, MD (Instituto Nacional de Ciencias Neurológicas, Lima, Perú); Manuela Verástegui, PhD; Javier Bustos, MD, MPH; Holger Mayta, PhD; Cristina Guerra, PhD; Yesenia Castillo, MSc; Yagahira Castro, MSc (Universidad Peruana Cayetano Heredia, Lima, Perú); María T. López, DVM, PhD; César M. Gavidia, DVM, PhD (School of Veterinary Medicine, Universidad Nacional Mayor de San Marcos, Lima, Perú); Luz M. Moyano, MD; Viterbo Ayvar, DVM (Cysticercosis Elimination Program, Tumbes, Perú); John Noh, BS and Sukwan Handali, MD (CDC, Atlanta, GA); Jon Friedland (Imperial College, London, UK).

Author Contributions

Conceived and designed the experiments: MJP HHG RHG SP MZ. Performed the experiments: MJP ME PS. Analyzed the data: MJP ME AG SM TN HHG RHG SP MZ. Contributed reagents/materials/analysis tools: PS AG PW HHG RHG MZ. Wrote the paper: MJP ME DR FG VC SM PW TN AG HHG RHG SP MZ. Sequenced and assembled the genomes: ED DR MR MF SS VC SA DB. Identified microsatellites: FG ME MJP MZ.

References

1. Coyle CM, Mahanty S, Zunt JR, Wallin MT, Cantey PT, White AC Jr, et al. Neurocysticercosis: neglected but not forgotten. PLoS Negl Trop Dis. 2012; 6: e1500. doi: 10.1371/journal.pntd.0001500 PMID: 22666505
2. Ndimubanzi PC, Carabin H, Budke CM, Nguyen H, Qian YJ, Rainwater E, et al. A systematic review of the frequency of neurocyticercosis with a focus on people with epilepsy. PLoS Negl Trop Dis. 2010; 4: e870. doi: 10.1371/journal.pntd.0000870 PMID: 21072231
3. Flisser A. Taeniasis and cysticercosis due to *Taenia solium*. Prog Clin Parasitol. 1994; 4: 77–116. PMID: 7948938
4. Nakao M, Okamoto M, Sako Y, Yamasaki H, Nakaya K, Ito A. A phylogenetic hypothesis for the distribution of two genotypes of the pig tapeworm *Taenia solium* worldwide. Parasitology. 2002; 124: 657–662. PMID: 12118722
5. Vega R, Pinero D, Ramanankandrasana B, Dumas M, Bouteille B, Fleury A, et al. Population genetic structure of *Taenia solium* from Madagascar and Mexico: implications for clinical profile diversity and immunological technology. Int J Parasitol. 2003; 33: 1479–1485. PMID: 14572511
6. Shrivastava J, Qian BZ, Mcvean G, Webster JP. An insight into the genetic variation of *Schistosoma japonicum* in mainland China using DNA microsatellite markers. Mol Ecol. 2005; 14: 839–849. PMID: 15723675
7. Campbell G, Garcia HH, Nakao M, Ito A, Craig PS. Genetic variation in *Taenia solium*. Parasitol Int. 2006; 55 Suppl: S121–6. PMID: 16352464
8. Barker GC. Microsatellite DNA: a tool for population genetic analysis. Trans R Soc Trop Med Hyg. 2002; 96 Suppl 1: S21–4. PMID: 12055841
9. Ito A, Yamasaki H, Nakao M, Sako Y, Okamoto M, Sato MO, et al. Multiple genotypes of *Taenia solium* —ramifications for diagnosis, treatment and control. Acta Trop. 2003; 87: 95–101. PMID: 12781383
10. Gasser RB, Zhu X, Woods W. Genotyping *Taenia* tapeworms by single-strand conformation polymorphism of mitochondrial DNA. Electrophoresis. 1999; 20: 2834–2837. PMID: 10546815
11. Mayta H, Talley A, Gilman RH, Jimenez J, Verastegui M, Ruiz M, et al. Differentiating *Taenia solium* and *Taenia saginata* infections by simple hematoxylin-eosin staining and PCR-restriction enzyme analysis. J Clin Microbiol. 2000; 38: 133–137. PMID: 10618076
12. Hancock K, Broughel DE, Moura IN, Khan A, Pieniazek NJ, Gonzalez AE, et al. Sequence variation in the cytochrome oxidase I, internal transcribed spacer 1, and Ts14 diagnostic antigen sequences of *Taenia solium* isolates from South and Central America, India, and Asia. Int J Parasitol. 2001; 31: 1601–1607. PMID: 11730787
13. Maravilla P, Souza V, Valera A, Romero-Valdovinos M, Lopez-Vidal Y, Dominguez-Alpizar JL, et al. Detection of genetic variation in *Taenia solium*. J Parasitol. 2003; 89: 1250–1254. PMID: 14740922
14. Maravilla P, Gonzalez-Guzman R, Zuniga G, Peniche A, Dominguez-Alpizar JL, Reyes-Montes R, et al. Genetic polymorphism in *Taenia solium* cysticerci recovered from experimental infections in pigs. Infect Genet Evol. 2008; 8: 213–216. doi: 10.1016/j.meegid.2007.11.006 PMID: 18243817
15. Bobes RJ, Fragoso G, Reyes-Montes Mdel R, Duarte-Escalante E, Vega R, de Aluja AS, et al. Genetic diversity of *Taenia solium* cysticerci from naturally infected pigs of central Mexico. Vet Parasitol. 2010; 168: 130–135. doi: 10.1016/j.vetpar.2009.11.001 PMID: 19963321
16. Speijer H, Savelkoul PH, Bonten MJ, Stobberingh EE, Tjhie JH. Application of different genotyping methods for *Pseudomonas aeruginosa* in a setting of endemicity in an intensive care unit. J Clin Microbiol. 1999; 37: 3654–3661. PMID: 10523569
17. Wassenaar TM, Newell DG. Genotyping of *Campylobacter* spp. Appl Environ Microbiol. 2000; 66: 1–9. PMID: 10618195
18. Schlotterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992; 20: 211–215. PMID: 1741246
19. Schlotterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma. 2000; 109: 365–371. PMID: 11072791
20. Madesis P, Ganopoulos I, Tsaftaris A. Microsatellites: evolution and contribution. Methods Mol Biol. 2013; 1006: 1–13. doi: 10.1007/978-1-62703-389-3_1 PMID: 23546780
21. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 2007; 25: 490–498. PMID: 17945369
22. Novelli VM, Cristofani-Yaly M, Bastianel M, Palmieri DA, Machado MA. Screening of genomic libraries. Methods Mol Biol. 2013; 1006: 17–24. doi: 10.1007/978-1-62703-389-3_2 PMID: 23546781

23. Oura CA, Odongo DO, Lubega GW, Spooner PR, Tait A, Bishop RP. A panel of microsatellite and minisatellite markers for the characterisation of field isolates of *Theileria parva*. Int J Parasitol. 2003; 33: 1641–1653. PMID: 14636680

24. Russell R, Iribar MP, Lambson B, Brewster S, Blackwell JM, Dye C, et al. Intra and inter-specific microsatellite variation in the *Leishmania subgenus Viannia*. Mol Biochem Parasitol. 1999; 103: 71–77. PMID: 10514082

25. Oliveira RP, Broude NE, Macedo AM, Cantor CR, Smith CL, Pena SD. Probing the genetic population structure of *Trypanosoma cruzi* with polymorphic microsatellites. Proc Natl Acad Sci U S A. 1998; 95: 3776–3780. PMID: 9520443

26. Su X, Wellems TE. Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats. Genomics. 1996; 33: 430–444. PMID: 8661002

27. Nakao M, Sako Y, Ito A. Isolation of polymorphic microsatellite loci from the tapeworm *Echinococcus multilocularis*. Infect Genet Evol. 2003; 3: 159–163. PMID: 14522179

28. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, et al. The genomes of four tapeworm species reveal adaptations to parasitism. Nature. 2013; 496: 57–63. doi: 10.1038/nature12031 PMID: 23485966

29. Dean DA, Wadl PA, Hadziabdic D, Wang X, Trigiano RN. Analyzing microsatellites using the QIAxcel system. Methods Mol Biol. 2013; 1006: 223–243. doi: 10.1007/978-1-62703-389-3_16 PMID: 23546795

30. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821–829. doi: 10.1101/gr.074492.107 PMID: 18349386

31. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res. 2000; 10: 62–71. PMID: 10645951

32. Rozenand S, Skaletsky H. Primer3 on the WWW for General Users and for Biologist Programmers. In: Krawets S, Misener S, editors. Bioinformatics Methods and Protocols. Totowa, NJ: Humana Press; 2000. pp. 365–386.

33. Xiao N, Remais J, Brindley PJ, Qiu D, Spear R, Lei Y, et al. Polymorphic microsatellites in the human bloodfluke, *Schistosoma japonicum*, identified using a genomic resource. Parasit Vectors. 2011; 4: 13-3305-4-13.

34. Gasser RB, Chilton NB. Characterisation of taeniid cestode species by PCR-RFLP of ITS2 ribosomal DNA. Acta Trop. 1995; 59: 31–40. PMID: 7785524

35. Castagnone-Sereno P, Danchin EG, Deleury E, Guillemaud T, Malausa T, Abad P. Genome-wide survey and analysis of microsatellites in nematodes, with a focus on the plant-parasitic species Meloidogyne incognita. BMC Genomics. 2010; 11: 598-2164-11-598.

36. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 2000; 10: 72–80. PMID: 10645952

37. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 2002; 11: 2453–2465. PMID: 12453231

38. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa L.*): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 2001; 11: 1441–1452. PMID: 11483586

39. Grover A, Sharma PC. Microsatellite motifs with moderate GC content are clustered around genes on *Arabidopsis thaliana* chromosome 2. In Silico Biol. 2007; 7: 201–213. PMID: 17688446

40. Dokholyan NV, Buldyrev SV, Havlin S, Stanley HE. Distributions of dimeric tandem repeats in non-coding and coding DNA sequences. J Theor Biol. 2000; 202: 273–282. PMID: 10666360

41. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 2000; 10: 967–981. PMID: 10899146

42. Jeri C, Gilman RH, Lescano AG, Mayta H, Ramirez ME, Gonzalez AE, et al. Species identification after treatment for human taeniasis. Lancet. 2004; 363: 949–950. PMID: 15043964

43. Yanagida T, Carod JF, Sako Y, Nakao M, Hoberg EP, Ito A. Genetics of the pig tapeworm in madagascar reveal a history of human dispersal and colonization. PLoS One. 2014; 9: e109002. doi: 10.1371/journal.pone.0109002 PMID: 25329310